UNIVERSITY OF CALIFORNIA
SANTA CRUZ

BAYESIAN MODELING AND INFERENCE FOR QUANTILE MIXTURE REGRESSION

A dissertation submitted in partial satisfaction of the
requirements for the degree of

DOCTOR OF PHILOSOPHY

in

STATISTICS AND APPLIED MATHEMATICS

by

Yifei Yan

September 2017

The Dissertation of Yifei Yan
is approved:

Professor Athanasios Kottas, Chair

Professor Juhee Lee

Professor Bruno Sansó

Tyrus Miller
Vice Provost and Dean of Graduate Studies

Table of Contents

List of Figures

List of Tables

Abstract

Bayesian Modeling and Inference for Quantile Mixture Regression

by

Yifei Yan

The focus of this work is to develop a Bayesian framework to combine information from multiple parts of the response distribution characterized with different quantiles. The goal is to obtain a synthesized estimate of the covariate effects on the response variable as well as to identify the more influential predictors. This framework naturally relates to the traditional quantile regression, which studies the relationship between the covariates and the conditional quantile of the response variable and serves as an attractive alternative to the more widely used mean regression methods. We achieve the objectives through constructing a Bayesian mixture model using quantile regressions as the mixture components.

The first stage of the research involves the development of a parametric family of distributions to provide the mixture kernel for the Bayesian quantile mixture regression. We derive a new family of error distributions for model-based quantile regression called generalized asymmetric Laplace distribution, which is constructed through a structured mixture of normal distributions. The construction enables fixing specific percentiles of the distribution while, at the same time, allowing for varying mode, skewness and tail behavior. This family provides a practically important extension of the asymmetric Laplace distribution, which is the standard error distribution for parametric quantile regression. We develop a Bayesian formulation for the proposed quantile regression model, including conditional lasso regularized quantile regression based on a hierarchical Laplace prior for the regression coefficients, and a Tobit quantile regression model.

Next, we develop the main framework to model the conditional distribution of the response with a weighted mixture of quantile regression components. We specify a common regression coefficient vector for all components to synthesize information from multiple parts of the response distribution, each modeled with one quantile regression component. The goal is to obtain a combined estimate of the predictive effect of each covariate. We consider the following two choices of kernel densities for the mixture model. When the probability of the quantile in each regression component is known, we model the components with the generalized asymmetric Laplace distribution, as its shape parameter introduces flexibility in shape and skewness to the kernel; else when the quantile probabilities are unknown, we use the asymmetric Laplace distribution as kernel density and view its skewness parameter, which is also the quantile probability of the component, as a random quantity and estimate it from the data. Under each kernel density, we formulate the hierarchical structure of the mixture weights and develop the approach to the posterior inference. We consider both parametric and nonparametric priors for the framework, and explore inferences for the number of components to be included. We demonstrate the performance of the method in identification of influential variables with simulation examples and illustrate the posterior predictive inferences in a realty price data from the Boston metropolitan area.

Finally, we extend the framework to apply the methods to specific problems in survival analysis and epidemiology. Both applications involve analyses of two cohorts, which oftentimes exhibit differing responses given the same predictor input. We adapt the proposed framework to model the survival data with right-censoring. For applications in epidemiology, we study the ordering properties of the mixture kernels and incorporate stochastic ordering in the two-cohort mixture framework through structured priors, which conforms with the assumption in certain circumstances of receiver operating characteristic curve estimation. With the adapted models, we carry out cohort-specific identification of influential variables and gain insights into the contribution in estimation and prediction from different parts of the response distribution, which are depicted by the corresponding quantile regression components. We illustrate the applications with a time-to-event data set on length of stay at nursing home and two disease diagnosis data sets, one on adolescent depression and the other on cattle epidemics.

To my parents and my best friends

Acknowledgments

I would like to extend my most sincere gratitude to my advisor, Professor Athanasios Kottas, who has always been accessible, patient, humorous, insightful and supportive. When I am faced with challenging problems and feel frustrated, Thanasis always encourages me and offers constructive advice and guidance; when I make progress, he shares my happiness. The past three years of dissertation research has been truly enjoyable, valuable and unforgettable. I would always remember the discussions we had about the research, as well as those outside of research, like on soccer games, "boring" movies by Andrei Tarkovsky and Monty Python's "Always Look on the Bright Side of Life".

I would like to acknowledge Professor Bruno Sansó and Professor Juhee Lee for taking the time to serve on my committee, reading my proposal and dissertation, and providing enlightening feedback and suggestions. My gratitude extends to my peers and professors in the AMS department, who offered tremendous support in my academic and personal life. I would like to thank my dear friend, Dr. Qinhua Zhou, who shared with me her personal experiences of tackling research challenges, encouraging me throughout.

My deepest gratitude goes to my beloved parents and my best friends in China. Although living on the other side of the Pacific, my parents have been constantly supportive and they always encourage me to trust myself, have faith in my decisions and pursue my dreams. I owe everything to them and I would not be here today without their love, sacrifice and support. My best friends, Le Wenyi, Shi Lingyi and Shi Xiaojing, have always been by my side even if they are physically not. They cheer me up when I feel disheartened. They fill me with love and courage and keep me going.

Finally, I thank life and circumstances for bringing me to Santa Cruz, to embark on this journey and complete it.

Chapter 1

Introduction

1.1 Background on quantile regression

Quantile regression studies the relation between the quantiles of the response variable and a given set of covariates. Unlike the usual mean regression that focuses on the expectation of the response variable, quantile regression not only allows for analysis of the center of the distribution through the median, but also applies to essentially any quantile of the distribution, such as the quartiles and the 5th and the 95th percentile that lie in the tails. Since the seminal work of Koenker & Bassett Jr (1978), quantile regression has seen extensive applications in various fields, such as ecology, epidemiology, sociology, economics and finance (Reich, 2012; Lee & Neocleous, 2010; Lum et al., 2012; Hu et al., 2013; Taddy & Kottas, 2010). It is a useful regression tool especially for problems wherein no relationship or only a weak relationship exists between the covariates and the conditional mean of the response variable (Cade & Noon, 2003).

As a practical and important alternative to traditional mean regression, quantile regression forms an area with a rapidly increasing literature. Parametric quantile regression models are almost exclusively built from the asymmetric Laplace (AL) distribution, the density of which is

$$f_p^{\mathrm{AL}}(y \mid \mu, \sigma) \quad = \quad \frac{p(1-p)}{\sigma} \exp\left\{-\frac{1}{\sigma}\rho_p\left(y-\mu\right)\right\}, \quad y \in \mathbb{R}$$

where $\rho_p(u) = u[p - I(u < 0)]$, with $I(\cdot)$ denoting the indicator function. Here, $\sigma > 0$ is a scale parameter, $\mu \in \mathbb{R}$ corresponds to the pth quantile, and $p \in (0,1)$ is the percentage, or cumulatively probability at the quantile μ, such that $\int_{-\infty}^{\mu} f_p^{\mathrm{AL}}(y \mid \mu, \sigma)\mathrm{d}y = p$. Hence, a model for pth quantile regression can be developed by expressing μ as a function of available covariates $\boldsymbol{x}$, for instance, $\mu = \boldsymbol{x}^T\boldsymbol{\beta}$ yields a linear quantile regression structure.

An important reason for its popularity in quantile regression is that maximizing the likelihood with respect to $\boldsymbol{\beta}$ under an AL response distribution corresponds to minimizing for $\boldsymbol{\beta}$ the check loss function, $\sum_{i=1}^{n} \rho_p(y_i - \boldsymbol{x}_i^T\boldsymbol{\beta})$, which is used for classical semi-parametric quantile regression (Koenker, 2005). Bayesian inference for quantile regression under AL errors is discussed in Yu & Moyeed (2001), Tsionas (2003) and Kozumi & Kobayashi (2011). Examples of applications of AL-based Bayesian quantile regression include analysis of repeated measure clinical trial data (Geraci & Bottai, 2007) and risk factor assessment for violent crime rate (Wang & Zhang, 2012).

However, if viewed as an error model for quantile regression, the AL distribution has substantial limitations. Most striking is that the skewness of the error density is fully determined when a specific percentile is chosen, that is, when p is fixed. In particular, the error density is symmetric in the case of median regression, since for $p = 0.5$, the AL reduces to the Laplace distribution. Moreover, the mode of the error distribution is at zero, for any p, which results in rigid error density tails for extreme percentiles. Yu & Zhang (2005) proposed a four-parameter variation of the AL distribution as a generalization, which however does not overcome either of the above issues of rigidity.

Given the limitations of the AL distribution, Bayesian semi-parametric and nonparametric approaches have been proposed for more flexible quantile regression in the past few years. The literature includes Bayesian nonparametric models for the error distribution in the special case of median regression (Walker & Mallick, 1999; Kottas & Gelfand, 2001; Hjort & Petrone, 2007). As for general quantile regression, Hjort & Petrone (2007) introduced a method based on the definition of Dirichlet process (DP) and Hjort & Walker (2009) adopted the idea of Pólya tree to develop a quantile pyramid process that supports piecewise linear quantile functions. Kottas & Krnjajić (2009) constructed a semi-parametric framework through scale DP mixtures of uniform densities. The posterior error densities have flexible skewness, yet are discontinuous at 0 resulting from fixing p. Reich et al. (2010) constructed a DP mixture of specifically designed normal mixtures estimated with a series of Metropolis-Hastings steps. Thompson et al. (2010) considered modeling quantiles of the covariates with natural cubic splines and the posterior sampling also involves a specially tuned Metropolis-Hastings algorithm.

There is limited work on parametric alternatives to AL quantile regression errors and the existing models do not overcome the major limitations discussed above. For instance, Wichitaksorn et al. (2014) studied a class of skew distributions for Bayesian quantile regression, but similar to the AL distribution, both the skewness and the percentile are controlled by the same parameter. Alternatively, Zhu & Zinde-Walsh (2009) and Zhu &

Galbraith (2011) explored the family of asymmetric exponential power distributions. Although it allows for different decay rates in the left and the right tails for fixed p, the mode of the distribution is held fixed at the quantile μ by construction, which constrains the behavior of the density around μ. Noufaily & Jones (2013) developed a parametric quantile regression based on generalized Gamma distributions, which covers a good range of shapes of distributions. However, since generalized gamma distributions are defined on $\mathbb{R}^+$, the method only applies to positive response variables.

On the other hand, in addition to regression of a single quantile, different Bayesian models were also developed for estimation of multiple quantiles with regression analysis, known as simultaneous quantile regression. In this context, each quantile is estimated with a different set of regression coefficients. Scaccia & Green (2003) modeled the conditional distribution of the response given a single continuous covariate with a discrete normal mixture with covariate-dependent weights. Further, Taddy & Kottas (2010) modeled the joint distribution of the response and the covariates with a DP mixture and developed inference for different quantile curves based on the induced conditional distribution of the response given the covariates. Tokdar & Kadane (2011) developed a semi-parametric model for simultaneous regression of multiple quantiles that satisfies the monotonicity constraint through an interpolation of two monotone curves. Reich & Smith (2013) proposed a Bayesian simultaneous quantile regression model for censored survival data with specifically designed basis functions. Das & Ghosal (2017) also considered basis function and represented the quantile function with a convex combination of two sets of B-spline basis expansions.

1.2 Motivation and objectives

Standard quantile regression focuses on a particular quantile of interest or on estimating a set of quantiles separately. We would like to go beyond the idea of individual quantile analysis and develop a Bayesian quantile mixture regression (BQMR) framework to combine information from multiple parts of the response distribution for the estimation of the regression coefficients.

More specifically, we propose to model the response distribution with a weighted mixture of K quantile regressions, each parameterized in terms of the p_kth quantile, $k = 1, \ldots, K$, where the p_k are ordered (for instance, $\{p_k\}$ can be equally spaced on the unit interval). By employing a *common* vector of regression coefficients $\boldsymbol{\beta}$ in all components, we obtain a combined estimate for the covariate effects. Such mixture framework has a great potential in identifying influential predictors, because instead of focusing on the mean or a single quantile of the response variable, it attends to multiple parts of the response distribution. If a predictor has differential effects on the higher and the lower responses, its regression coefficient will be attenuated when the mixture model synthesizes these effects in the estimation. Combined with sparsity-inducing priors such as Bayesian lasso (Park & Casella, 2008), the proposed framework will naturally select the predictors that have a consistent effect across different parts of the response distribution.

Our proposed framework may remind readers of the composite quantile regression (CQR) in the classical literature (Zou & Yuan, 2008), which is a composite analysis also on multiple quantiles. Although the two approaches have similar objectives, there exists a clear distinction. The CQR procedure focuses on the check loss of K equally weighted quantile regressions with a common regression coefficient vector $\hat{\boldsymbol{\beta}}^{CQR}$ through minimizing the total check loss summed up across all K quantiles. Zou & Yuan (2008) shows that CQR is a powerful variable selection tool, as the estimator enjoys the oracle properties under adaptive lasso penalty. However, devised to minimize the check loss, the CQR procedure is purely optimization-oriented and does not involve probabilistic modeling. The Bayesian framework we propose targets the response distribution directly through a mixture model and the inference follows naturally from a modeling perspective.

We would like to also emphasize the differences between the proposed mixture framework and simultaneous quantile regression (Taddy & Kottas, 2010; Tokdar & Kadane, 2011; Reich & Smith, 2013; Das & Ghosal, 2017), which models several quantiles jointly. Unlike what we propose in this work, simultaneous quantile regression does not involve a mixture modeling framework. Quite the contrary, it handles each individual quantile in a separate regression: the p_kth quantile has its own vector of regression coefficients $\boldsymbol{\beta}_{p_k}$, which describes the covariate effect specific to the quantile. What we attempt to achieve in the mixture model is to assess the comprehensive effect of the covariates on the response distribution. This is realized through estimating a common regression coefficient $\boldsymbol{\beta}$ shared by multiple quantile regression components.

To construct the proposed Bayesian mixture quantile regression model, the first step is to develop a new parametric distribution that is more general than the AL distribution as the kernel of mixture components. Chapter 2 presents the development, implementation and applications of such distribution. Given the limitations of the AL distribution, we believe that developing a more flexible parametric error distribution for quantile analysis is of independent interest. We derive the new distribution through constructively modifying the mixture representation of the AL distribution. Through introducing a shape parameter, we obtain a family of distributions that has more flexible skewness and tail behaviour than the AL, while including it as a special case of the family. The resulting distribution has flexible skewness and mode as well as a continuous density function. We develop a Bayesian formulation for the proposed quantile regression model, including conditional lasso regularized quantile regression based on a hierarchical Laplace prior for the regression coefficients and a Tobit quantile regression model. The Markov chain Monte Carlo (MCMC) algorithm under the new distribution sacrifices very little in terms of ease of implementation, since save for one parameter, it is based on all Gibbs updates.

Next, we construct the Bayesian quantile mixture regression (BQMR) framework and demonstrate its advantages in prediction and identification of influential variables, particularly in the cases where the errors follow a nonstandard distribution with heavy-skewness or multi-modality in the density function. We develop two versions of the BQMR framework under different specifications of kernel density for the mixture. Constructed with the proposed new distribution, the first version of the framework is designed to study the important predictors that affect specific areas of the response distribution. The second version of the framework is based upon mixtures of AL distributions. To overcome the aforementioned limitations of the AL distribution, we treat the probability parameter p as a random variable and estimate it from the data. This framework is tailored to obtain inferences on the percentage $\{p_k\}$ of the components and to explore the response distribution in an efficient manner. Further, under both versions of BQMR, visualization of the posterior predictive inference of the mixture components offers insight to how different quantile components contribute to the analysis. Illustrations of the models indicate that they offer interesting insights to the conditional distribution of the response variable, in addition to identifying influential variables via integrating information from multiple parts of the response distribution. Construction, implementation and related extensions of the BQMR framework are presented in Chapter 3.

Finally in the Chapter 4, we study the application of the BQMR framework in survival analysis and estimation of receiver operating characteristic (ROC) curve, both of which are important topics in biomedical research. The applications we consider involve data from two cohorts in controlled studies. We explore extensions of the framework to handle censored observations and develop predictive inference for the survival functions and ROC curves while allowing for cohort-specific identification of influential variables. We further derive theoretical results on stochastic ordering with the proposed BQMR. Based on the findings, we develop a two-cohort BQMR framework. The framework models the response from both cohorts at the same time and ensures stochastic ordering of the two cohorts by construction, which is achieved through careful specification of the priors.

Chapter 2

A new family of error distributions for Bayesian quantile regression

The focus of this chapter is the development of a new family of error distributions for Bayesian quantile regression. First, we construct the new distribution and discuss its properties relative to the AL distribution in Section 2.1. We then formulate the Bayesian quantile regression model in Section 2.2, including a prior specification for the regression coefficients that encourages shrinkage resulting in regularized quantile regression, and a Tobit quantile regression formulation. In Section 2.3, we present results from a simulation study to compare the performance of the AL and the proposed distribution in regularized quantile regression. The methodology is illustrated with three data examples in Section 2.4, focusing again on comparison with the AL quantile regression model.

2.1 Theory and properties of the new error distribution

In this section, we derive the distribution and explore some key properties of the new family.

2.1.1 The generalized asymmetric Laplace distribution

The construction of the new distribution is motivated by the most commonly used mixture representation of the AL density. In particular,

$$f_p^{\mathrm{AL}}(y \mid \mu, \sigma) = \int_{\mathbb{R}^+} \mathrm{N}(y \mid \mu + \sigma A(p) z, \sigma^2 B(p) z) \, \mathrm{Exp}(z \mid 1) \, \mathrm{d}z \tag{2.1}$$

where $A(p) = (1-2p)/\{p(1-p)\}$ and $B(p) = 2/\{p(1-p)\}$. Moreover, $\mathrm{N}(m, W)$ denotes the normal distribution with mean m and variance W, and $\mathrm{Exp}(1)$ denotes the exponential distribution with mean 1. We use such notation throughout to indicate either the distribution or its density, depending on the context.

The mixture formulation in (2.1) enables exploration of extensions to the AL distribution. Extending the $\mathrm{Exp}(1)$ mixing distribution is not a fruitful direction in terms of evaluation of the integral, and, more importantly, with respect to fixing percentiles of the resulting distribution. However, both goals are accomplished by replacing the normal kernel in (2.1) with a skew normal kernel (Azzalini, 1985). In its original parameterization, the skew normal density is given by $f^{\mathrm{SN}}(y \mid \xi, \omega, \lambda) = 2\omega^{-1} \, \phi(\omega^{-1}(y-\xi)) \, \Phi(\lambda\omega^{-1}(y-\xi))$, where $\phi(\cdot)$ and $\Phi(\cdot)$ denote the density and distribution function, respectively, of the standard normal distribution. Here, $\xi \in \mathbb{R}$ is a location parameter, $\omega > 0$ a scale parameter, and $\lambda \in \mathbb{R}$ the skewness parameter. Key to our construction is the fact that the skew normal density can be written as a location normal mixture with mixing distribution given by a standard normal truncated on $\mathbb{R}^+$ (Henze, 1986). We reparameterize (ξ, ω, λ) to (ξ, τ, ψ), where $\tau > 0$ and $\psi \in \mathbb{R}$, such that $\lambda = \psi/\tau$ and $\omega = (\tau^2 + \psi^2)^{1/2}$. Then the density can be written as $f^{\mathrm{SN}}(y \mid \xi, \tau, \psi) = \int_{\mathbb{R}^+} \mathrm{N}(y \mid \xi + \psi s, \tau^2)\mathrm{N}^+(s \mid 0, 1)\,\mathrm{d}s$, where $\mathrm{N}^+(0, 1)$ denotes the standard normal distribution truncated over $\mathbb{R}^+$.

The proposed model, referred to as generalized asymmetric Laplace (GAL) distribution, is built by adding a shape parameter, $\alpha \in \mathbb{R}$, to the mean of the normal kernel in (2.1) and mixing with respect to a $\mathrm{N}^+(0, 1)$ variable. More specifically, the full mixture representation for the density function, $f(y \mid p, \alpha, \mu, \sigma)$, of the new distribution is as follows

$$\iint_{\mathbb{R}^+ \times \mathbb{R}^+} \mathrm{N}(y \mid \mu + \sigma\alpha s + \sigma A(p)z, \sigma^2 B(p)z)\,\mathrm{Exp}(z \mid 1)\,\mathrm{N}^+(s \mid 0, 1)\,\mathrm{d}z\mathrm{d}s. \tag{2.2}$$

Note that, integrating over s in (2.2), the GAL density can be expressed in the form of (2.1) with the $\mathrm{N}(y \mid \mu + \sigma A(p)z, \sigma^2 B(p)z)$ kernel replaced with a skew normal kernel, which, in its original parameterization, has location parameter $\mu + \sigma A(p)z$, scale parameter $\sigma\{\alpha^2 + B(p)z\}^{1/2}$, and skewness parameter $\alpha\{B(p)z\}^{-1/2}$. Evidently, when $\alpha = 0$, $f(y \mid p, 0, \mu, \sigma)$ reduces to the AL density.

To obtain the GAL density, we integrate out first z and then s in (2.2). The integrand of $\int_{\mathbb{R}^+} \mathrm{N}(y \mid \mu + \sigma\alpha s + \sigma A(p)z, \sigma^2 B(p)z)\,\mathrm{Exp}(z \mid 1)\,\mathrm{d}z$ can be recognized as the kernel of a generalized inverse-Gaussian density. Integrating out z, we obtain $f(y \mid p, \alpha, \mu, \sigma) = \int_{\mathbb{R}^+} p(1-p)\sigma^{-1} \exp\left\{-\sigma^{-1}\left[p - I(y < \mu + \sigma\alpha s)\right]\left[y - (\mu + \sigma\alpha s)\right]\right\} \mathrm{N}^+(s \mid 0, 1)\,\mathrm{d}s$. This integral involves a normal density kernel, but care is needed with the limits of integration which depend on the sign of $y - \mu$ and of α. Combining the resulting expressions from all possible cases, we obtain that for $\alpha \neq 0$, the GAL density $f(y \mid p, \alpha, \mu, \sigma)$ is given by

$$2\,\frac{p(1-p)}{\sigma}\left(\left[\Phi\left(\frac{y^*}{\alpha} - p_{\alpha_-}\alpha\right) - \Phi(-p_{\alpha_-}\alpha)\right]\exp\left\{-p_{\alpha_-}y^* + \frac{1}{2}(p_{\alpha_-}\alpha)^2\right\} I\left(\frac{y^*}{\alpha} > 0\right)\right.$$

$$+\,\Phi\left[p_{\alpha_+}\alpha-\frac{y^*}{\alpha}I\left(\frac{y^*}{\alpha}>0\right)\right]\exp\left\{-p_{\alpha_+}y^*+\frac{1}{2}\left(p_{\alpha_+}\alpha\right)^2\right\}\Bigg) \tag{2.3}$$

where $y^* = (y - \mu)/\sigma$, $p_{\alpha_+} = p - I(\alpha > 0)$, $p_{\alpha_-} = p - I(\alpha < 0)$, with $p \in (0, 1)$. The relatively complex form of the density in (2.3) is not an obstacle from a practical perspective, since its hierarchical mixture representation facilitates study of model properties and Markov chain Monte Carlo posterior simulation.

2.1.2 Link to the p_0th quantile

There is a direct link between the GAL distribution and the p_0th quantile for any $p_0 \in (0, 1)$; note that parameter p no longer corresponds to the cumulative probability at the quantile for $\alpha \neq 0$. When $\alpha > 0$, the distribution function of (2.3) at μ is given by $\int_{-\infty}^{\mu} f(y \mid p, \alpha, \mu, \sigma)\mathrm{d}y = 2p\Phi[(p-1)\alpha]\exp\left\{(p-1)^2\alpha^2/2\right\}$. Hence, letting $\gamma = (1-p)\alpha$, the distribution function becomes,

$$\int_{-\infty}^{\mu} f(y \mid p, \gamma, \mu, \sigma)\,\mathrm{d}y = p\,g(\gamma) \qquad \text{with} \qquad g(\gamma) = 2\Phi(-|\gamma|)\exp(\gamma^2/2).$$

Because $p \in (0, 1)$ by definition, the reparameterization of $\gamma = (1-p)\alpha$ implies that γ and α have the same sign. We use $|\gamma|$ above, since this is the general form of $g(\gamma)$ that applies also in the $\alpha < 0$ case.

In the following, we show that $g(\gamma)$ is monotonically increasing in $\mathbb{R}^-$ and monotonically decreasing in $\mathbb{R}^+$, an important property that enables fixing the percentile of the distribution. For $\gamma \in \mathbb{R}^-$, $\mathrm{d}g(\gamma)/\mathrm{d}\gamma = 2h(\gamma)\exp(\gamma^2/2)$, where $h(\gamma) = \phi(\gamma) + \gamma\Phi(\gamma)$. The function $h(\gamma)$ is monotonically increasing in $\mathbb{R}^-$, since $\mathrm{d}h(\gamma)/\mathrm{d}\gamma = \Phi(\gamma) > 0$. Moreover, $h(0) = (2\pi)^{-1/2} > 0$, and $\lim_{\gamma\to-\infty} h(\gamma) = 0$. Therefore, $h(\gamma) > 0$ for $\gamma \in \mathbb{R}^-$, and thus

$g(\gamma)$ is monotonically increasing in $\mathbb{R}^-$. Since $g(\gamma)$ is an even function, it also obtains that it is monotonically decreasing in $\mathbb{R}^+$.

Consider now setting $\int_{-\infty}^{\mu} f(y \mid p, \gamma, \mu, \sigma)\,\mathrm{d}y = pg(\gamma) = p_0$. Recall that α and γ have the same sign. Then for each $\gamma > 0$ in the domain that respects the condition of $p \in (0, 1)$, there is a unique solution of p that ensures $\int_{-\infty}^{\mu} f(\epsilon|\cdot) = p_0$, and subsequently a unique α based on $\gamma = (1 - p)\alpha$. For $\alpha < 0$, setting $\int_{\mu}^{\infty} f(y \mid p, \gamma, \mu, \sigma)\,\mathrm{d}y = 1 - p_0$ and letting $\gamma = p\alpha$ leads to the same argument.

The above connection between (p_0, γ) and (p, α) suggests that by reparameterization with desired p_0 and $\gamma = [I(\alpha > 0) - p]|\alpha|$, we can derive a new family of distributions with the percentile for fixed p_0 given by μ, and with an additional shape parameter γ. For $\gamma \neq 0$, the density, $f_{p_0}(y \mid \gamma, \mu, \sigma)$, of such quantile-fixed GAL distribution is

$$2\,\frac{p(1-p)}{\sigma}\left(\left\{\Phi\left(-\frac{p_{\gamma_+}y^*}{|\gamma|} + \frac{p_{\gamma_-}}{p_{\gamma_+}}|\gamma|\right) - \Phi\left(\frac{p_{\gamma_-}}{p_{\gamma_+}}|\gamma|\right)\right\}\exp\left\{-p_{\gamma_-}y^* + \frac{\gamma^2}{2}\left(\frac{p_{\gamma_-}}{p_{\gamma_+}}\right)^2\right\} I\left(\frac{y^*}{\gamma} > 0\right)\right.$$
$$\left. + \Phi\left[-|\gamma| + \frac{p_{\gamma_+}y^*}{|\gamma|} I\left(\frac{y^*}{\gamma} > 0\right)\right]\exp\left\{-p_{\gamma_+}y^* + \frac{\gamma^2}{2}\right\}\right) \tag{2.4}$$

where $p \equiv p(\gamma, p_0) = I(\gamma < 0) + \{[p_0 - I(\gamma < 0)]/g(\gamma)\}$, $p_{\gamma_+} = p - I(\gamma > 0)$, $p_{\gamma_-} = p - I(\gamma < 0)$, and $y^* = (y - \mu)/\sigma$. Parameter γ has bounded support over interval (L, U), where L is the negative root of $g(\gamma) = 1 - p_0$ and U is the positive root of $g(\gamma) = p_0$. For instance, γ takes values in $(-0.07, 15.90)$, $(-1.09, 1.09)$ and $(-2.90, 0.39)$ when $p_0 = 0.05$, $.5$ and 0.75, respectively. When $\gamma = 0$, the density of GAL reduces to the AL density, which is also a limiting case of (2.4). The density function is continuous for all possible γ values.

Additionally, the cumulative distribution function (CDF) $F(y)$ of a quantile-fixed GAL distribution for the p_0th quantile can be obtained in closed form,

i) For $\gamma < 0$,

$$F(y) \quad = \quad \left\{ \begin{array}{ll} 2\Phi\left[-\frac{p(y-\mu)}{\gamma\sigma}\right] + 2p\left\{\Phi\left[\frac{p(y-\mu)}{\gamma\sigma} - \frac{(p-1)\gamma}{p}\right] - \Phi\left[-\frac{(p-1)\gamma}{p}\right]\right\} & \\ \quad \cdot\exp\left\{-\frac{(p-1)(y-\mu)}{\sigma} + \frac{(p-1)^2\gamma^2}{2p^2}\right\} - 2(1-p)\Phi\left[\gamma - \frac{p(y-\mu)}{\gamma\sigma}\right] & \\ \quad \cdot\exp\left\{-\frac{p(y-\mu)}{\sigma} + \frac{\gamma^2}{2}\right\} \ , & y \le \mu \\ p_0 + (1-p_0)\left[1 - \exp\left\{-\frac{p(y-\mu)}{\sigma}\right\}\right] \ , & y > \mu \end{array} \right.$$

ii) For $\gamma = 0$ (AL distribution),

$$F(y) \quad = \quad \left\{ \begin{array}{ll} p_0 \exp\left\{-(p_0 - 1)\frac{y-\mu}{\sigma}\right\} \ , & y \le \mu \\ p_0 + (1-p_0)\left[1 - \exp\left\{-\frac{p_0(y-\mu)}{\sigma}\right\}\right] \ , & y > \mu \end{array} \right.$$

iii) For $\gamma > 0$,

$$F(y) \quad = \quad \left\{ \begin{array}{ll} p_0 \exp\left\{-(p-1)\frac{y-\mu}{\sigma}\right\} \ , & y \le \mu \\ 2\Phi\left[\frac{(1-p)(y-\mu)}{\gamma\sigma}\right] - 1 - 2(1-p)\left\{\Phi\left[\frac{(1-p)(y-\mu)}{\gamma\sigma} - \frac{p\gamma}{1-p}\right] - \Phi\left(-\frac{p\gamma}{1-p}\right)\right\} & \\ \quad \cdot\exp\left\{-\frac{p(y-\mu)}{\sigma} + \frac{p^2\gamma^2}{2(1-p)^2}\right\} + 2p\,\Phi\left[-\gamma - \frac{(1-p)(y-\mu)}{\gamma\sigma}\right] & \\ \quad \cdot\exp\left\{-\frac{(p-1)(y-\mu)}{\sigma} + \frac{\gamma^2}{2}\right\} \ , & y > \mu \end{array} \right.$$

where p and $g(\gamma)$ are defined as are in the density function.

2.1.3 Properties of the GAL distribution

The quantile-fixed GAL distribution has three parameters, μ, σ and γ. Note that Y has density $f_{p_0}(\cdot \mid \gamma, \mu, \sigma)$ if and only if $(Y-\mu)/\sigma$ has density $f_{p_0}(\cdot \mid \gamma, 0, 1)$. Hence, similarly to the AL distribution, μ is a location parameter and σ is a scale parameter. The new shape parameter γ enables the extension relative to the quantile-fixed AL distribution. As demonstrated in Figure 2.1, γ controls skewness and tail behaviour, allowing for both

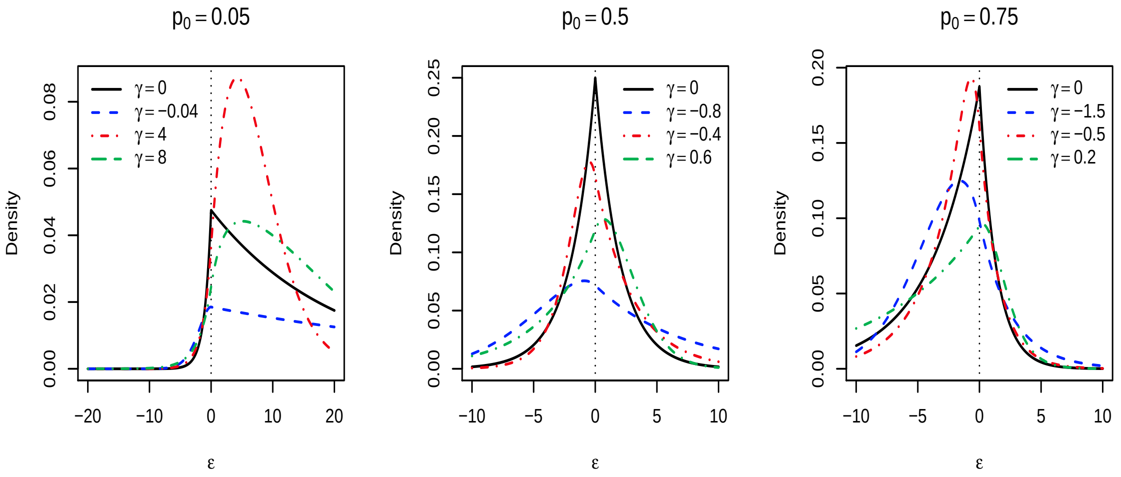


Figure 2.1: Density function of quantile-fixed generalized asymmetric Laplace distribution with $\mu = 0$, $\sigma = 1$ and different values of γ, for $p_0 = 0.05$, 0.5 and 0.75. In all cases, the solid line corresponds to the asymmetric Laplace density ($\gamma = 0$).

left and right skewness when the median is fixed, as well as for both heavier and lighter tails than the asymmetric Laplace, the difference being particularly emphatic for extreme percentiles. Moreover, as γ varies, the mode is no longer held fixed at μ; it is less than μ when $\gamma < 0$ and greater than μ when $\gamma > 0$. The above attributes render the proposed distribution substantially more flexible than the AL distribution.

Moreover, we note that parameter γ satisfies likelihood identifiability. Consider the location-scale standardized density, $f_{p_0}(\cdot \mid \gamma, 0, 1)$, which is effectively the model for the errors in quantile regression. Then, assume $f_{p_0}(y \mid \gamma_1, 0, 1) = f_{p_0}(y \mid \gamma_2, 0, 1)$, for all $y \in \mathbb{R}$. Given that parameter γ controls the mode of the density, this implies that γ_1 and γ_2 must have the same sign. Working with either of the two cases (that is, $\gamma_1 > 0$ and $\gamma_2 > 0$ or $\gamma_1 < 0$ and $\gamma_2 < 0$) in expression (2.4), we arrive at $g(\gamma_1) = g(\gamma_2)$, which, based on the monotonicity of function $g(\cdot)$, implies $\gamma_1 = \gamma_2$.

Finally, we derive the characteristic function of GAL_{p_0} using twice the double expectation theorem based on the hierarchical representation.

$$\varphi_Y(t) \;=\; E(e^{itY}) \;=\; E_Z[E_{Y|Z}(e^{itY})] \;=\; E_Z[E_{S|Z}[E_{Y|S,Z}(e^{itY})]]$$

Since S follows a truncated normal distribution on $\mathbb{R}^+$, also known as the half-normal distribution, $E_{Y|Z}(e^{itY})$ can be obtained with the characteristic function of half-normal,

$$\begin{aligned} E_{S|Z}[E_{Y|S,Z}(e^{itY})] &= E_{S|Z}\left[e^{it(\mu+\sigma C|\gamma|s+\sigma Az)-\frac{t^2}{2}\sigma^2 Bz}\right] \\ &= e^{it(\mu+\sigma Az)-\frac{t^2}{2}\sigma^2 Bz}\varphi_{S|Z}(\sigma C|\gamma|t) \\ &= e^{it(\mu+\sigma Az)-\frac{t^2}{2}\sigma^2 Bz}\cdot 2e^{-\frac{t^2}{2}(\sigma C\gamma)^2}[1-\Phi(-i\sigma C|\gamma|t)] \end{aligned}$$

Denote $c = 2e^{it\mu-\frac{t^2}{2}(\sigma C\gamma)^2}[1-\Phi(-i\sigma C|\gamma|t)]$. Then the characteristic function of GAL_{p_0} is,

$$\varphi_Y(t) \;=\; E_Z[ce^{(it\sigma A-\frac{t^2}{2}\sigma^2 B)z}] \;=\; \frac{2e^{it\mu-\frac{t^2}{2}(\sigma C\gamma)^2}[1-\Phi(-i\sigma C|\gamma|t)]}{1-it\sigma A+\frac{t^2}{2}\sigma^2 B}$$

2.2 Bayesian inference for quantile regression under GAL$_{p_0}$

The mixture representation of the proposed distribution allows for straightforward implementation of Bayesian quantile regression under the GAL$_{p_0}$ distribution. In this section we develop inference for the regression model as well as for its extensions in regularized regression and Tobit regression.

2.2.1 Regression model formulation

Consider continuous responses y_i and the associated covariate vectors $\boldsymbol{x}_i$, for $i = 1,\ldots,n$. The linear quantile regression model is set up as $y_i = \boldsymbol{x}_i^T\boldsymbol{\beta} + \epsilon_i$, where the ϵ_i arise independently from a quantile-fixed GAL distribution with $\int_{-\infty}^{0} f_{p_0}(\epsilon \mid \gamma, 0, \sigma)\mathrm{d}\epsilon = p_0$.

Owing to the mixture representation of the new distribution, the model for the data can be expressed hierarchically as follows

$$\begin{aligned} y_i \mid \boldsymbol{\beta}, \gamma, \sigma, z_i, s_i &\overset{ind.}{\sim} \mathrm{N}(y_i \mid \boldsymbol{x}_i^T \boldsymbol{\beta} + \sigma C |\gamma| s_i + \sigma A z_i, \sigma^2 B z_i), \;\; i = 1, ..., n \\ z_i, s_i &\overset{ind.}{\sim} \mathrm{Exp}(z_i \mid 1)\, \mathrm{N}^+(s_i \mid 0, 1), \;\; i = 1, ..., n \end{aligned} \tag{2.5}$$

where $C = [I(\gamma > 0) - p]^{-1}$, and A and B are the functions of p given in (2.1). Since p is a function of γ and p_0, A, B and C are all functions of parameter γ. The Bayesian model is completed with priors for $\boldsymbol{\beta}$, σ and γ. Here, we assume a normal prior $\mathrm{N}(\boldsymbol{m}_0, \Sigma_0)$ for $\boldsymbol{\beta}$ and an inverse-gamma prior $\mathrm{IG}(a_\sigma, b_\sigma)$ for σ, with mean $b_\sigma/(a_\sigma - 1)$ provided $a_\sigma > 1$. For any specified p_0, γ is defined over an interval (L, U) with fixed finite endpoints, and thus a natural prior for γ is given by a rescaled Beta distribution, with the uniform distribution available as a default choice.

The augmented posterior distribution, which includes the z_i and the s_i, can be explored via a Markov chain Monte Carlo algorithm based on Gibbs sampling updates for all parameters other than γ. As in Kozumi & Kobayashi (2011), we set $v_i = \sigma z_i$, $i = 1, \ldots, n$. Then, the posterior simulation method is based on the following updates.

1. Sample $\boldsymbol{\beta}$ from $\mathrm{N}(\boldsymbol{m}^*, \Sigma^*)$, with covariance matrix $\Sigma^* = [\Sigma_0^{-1} + \sum_{i=1}^n \boldsymbol{x}_i \boldsymbol{x}_i^T/(B\sigma v_i)]^{-1}$ and mean vector $\boldsymbol{m}^* = \Sigma^* \{\Sigma_0^{-1} \boldsymbol{m}_0 + \sum_{i=1}^n \boldsymbol{x}_i [y_i - (\sigma C|\gamma| s_i + A v_i)]/(B\sigma v_i)\}$.

2. For each $i = 1, ..., n$, sample v_i from a generalized inverse-Gaussian distribution, $\mathrm{GIG}(0.5, a_i, b_i)$, where $a_i = [y_i - (\boldsymbol{x}_i^T \boldsymbol{\beta} + \sigma C|\gamma| s_i)]^2/(B\sigma)$ and $b_i = 2/\sigma + A^2/(B\sigma)$, with density given by $\mathrm{GIG}(x \mid \nu, a, b) \propto x^{\nu - 1} \exp\{-0.5(a/x + bx)\}$.

3. For each $i = 1, ..., n$, sample s_i from a normal $\mathrm{N}(\mu_{s_i}, \sigma_{s_i}^2)$ distribution truncated on

$\mathbb{R}^+$, where $\sigma_{s_i}^2 = [(C\gamma)^2\sigma/(Bv_i) + 1]^{-1}$ and $\mu_{s_i} = \sigma_{s_i}^2 C|\gamma|[y_i - (\boldsymbol{x}_i^T\boldsymbol{\beta} + Av_i)]/(Bv_i)$.

4. Sample σ from a GIG(ν, c, d) distribution, where $\nu = -(a_\sigma + 1.5n)$, $c = 2b_\sigma + 2\sum_{i=1}^n v_i + \sum_{i=1}^n [y_i - (\boldsymbol{x}_i^T\boldsymbol{\beta} + Av_i)]^2/(Bv_i)$, and $d = \sum_{i=1}^n (C\gamma s_i)^2/(Bv_i)$.

5. Update γ with a Metropolis-Hastings step, using a normal proposal distribution on the logit scale over (L, U).

Based on the hierarchical model structure, the posterior predictive error density can be expressed as $p(\epsilon \mid \text{data}) = \int \mathrm{N}(\epsilon \mid \sigma C|\gamma|s + \sigma Az, \sigma^2 Bz)\,\mathrm{Exp}(z \mid 1)\,\mathrm{N}^+(s \mid 0, 1)\,\pi(\gamma, \sigma \mid \text{data})\,\mathrm{d}s\,\mathrm{d}z\,\mathrm{d}\gamma\,\mathrm{d}\sigma$, and thus estimated through Monte Carlo integration, using the posterior samples of (γ, σ).

2.2.2 Quantile regression with regularization

Since the GAL distribution is constructed through modifying the mixture representation of the AL distribution, it retains some of the interesting properties of the AL distribution. In particular, working with the hierarchical representation of the GAL distribution, we are able to retrieve an extended version of the check loss function which corresponds to asymmetric Laplace errors.

Consider the collapsed posterior distribution, $\pi(\boldsymbol{\beta}, \gamma, \sigma, s_1, ..., s_n \mid \text{data})$, that arises from (2.5) by marginalizing over the z_i. Then, the corresponding posterior full conditional for $\boldsymbol{\beta}$ can be expressed as

$$\pi(\boldsymbol{\beta} \mid \gamma, \sigma, s_1, ..., s_n, \text{data}) \propto \pi(\boldsymbol{\beta}) \exp\left\{-\frac{1}{\sigma}\sum_{i=1}^n \rho_p(y_i - \boldsymbol{x}_i^T\boldsymbol{\beta} - \sigma H(\gamma)s_i)\right\}$$

where $\pi(\boldsymbol{\beta})$ is the prior density for $\boldsymbol{\beta}$, $H(\gamma) = C|\gamma| = \gamma g(\gamma)/\{g(\gamma) - |p_0 - I(\gamma < 0)|\}$, and $p = I(\gamma < 0) + \{[p_0 - I(\gamma < 0)]/g(\gamma)\}$, with p_0 the probability associated with the specified quantile modeled through $\boldsymbol{x}_i^T\boldsymbol{\beta}$. Hence, ignoring the prior contribution, finding the mode of the posterior full conditional for $\boldsymbol{\beta}$ is equivalent to minimizing with respect to $\boldsymbol{\beta}$ the adjusted loss function $\sum_{i=1}^n \rho_p(y_i - \boldsymbol{x}_i^T\boldsymbol{\beta} - \sigma H(\gamma)s_i)$; note that in the special case with asymmetric Laplace errors, that is, for $\gamma = 0$, this reduces to the check loss function with $p = p_0$.

Based on the above structure, the positive-valued latent variables s_i can be viewed as response-specific weights that are adjusted by real-valued coefficient $H(\gamma)$, which is fully specified through the shape parameter γ. The result is the real-valued, response-specific terms $\sigma H(\gamma)s_i$, which reflect on the estimation of $\boldsymbol{\beta}$ the effect of outlying observations relative to the AL distribution. A promising direction to further explore this structure is in the context of variable selection guided by the shrinkage of covariate effects. For instance, Li et al. (2010) study connections between different versions of regularized quantile regression and different priors for $\boldsymbol{\beta}$, working with asymmetric Laplace errors. The main example is lasso regularized quantile regression, which can be connected to the Bayesian asymmetric Laplace error model through a hierarchical Laplace prior for $\boldsymbol{\beta}$. We consider this prior for Bayesian quantile regression with the proposed GAL distribution. The perspective we offer may be useful, since it can be used to explore regularization adjusting the loss function through the response distribution, in addition to the penalty term through the prior for the regression coefficients.

Here, we denote by $\boldsymbol{\beta}$ the d-dimensional vector of regression coefficients excluding the intercept β_0. Then, the Laplace conditional prior structure for $\boldsymbol{\beta}$ is given by

$$\pi(\boldsymbol{\beta} \mid \sigma, \lambda) \quad = \quad \prod_{k=1}^{d} \frac{\lambda}{2\sigma} \exp\left\{-\frac{\lambda}{\sigma}|\beta_k|\right\} \quad = \quad \prod_{k=1}^{d} \int_{\mathbb{R}^+} \frac{1}{\sqrt{2\pi\omega_k}} \exp\left\{-\frac{\beta_k^2}{2\omega_k}\right\} \frac{\eta^2}{2} \exp\left\{-\frac{\eta^2}{2}\omega_k\right\} \mathrm{d}\omega_k$$

The expression following the second equals sign utilizes the normal scale mixture representation for the Laplace distribution, which has been exploited for posterior simulation in the context of lasso mean regression (Park & Casella, 2008). Moreover, to facilitate Markov chain Monte Carlo sampling, we reparameterize in terms of $\eta = \lambda/\sigma$ and place a gamma prior on η^2. The lasso regularized version of model (2.5) is completed with a normal prior for β_0, and with the priors for the other parameters as given in Section 2.2.1. The posterior simulation algorithm is the same with the one described in Section 2.2.1 with the exception of the updates for the β_k, $k = 1, ..., d$, and for η^2. Using the mixture representation of the Laplace prior, each β_k can be sampled from a normal distribution, whereas η^2 has a gamma posterior full conditional distribution (details provided in Appendix A.1).

2.2.3 Tobit quantile regression

Tobit regression offers a modeling strategy for problems involving range constraints on the response variable (Amemiya, 1984). The standard Tobit regression model can be viewed in the context of censored regression where the responses are left censored at a threshold c; without loss of generality, we take $c = 0$. The responses can be written as $y_i = \max\{0, y_i^*\}$, where y_i are the observed values and y_i^* are latent if $y_i^* \leq 0$. In the context of quantile regression, Yu & Stander (2007) and Kozumi & Kobayashi (2011) applied the AL-based model to the latent responses y_i^*. Here, we consider the Tobit quantile regression setting with GAL errors.

Consider a data set of $n + k$ observations on covariates and associated responses $\boldsymbol{y} = (\boldsymbol{y}^{\text{o}}, \mathbf{0})$, where $\boldsymbol{y}^{\text{o}} = (y_1^{\text{o}}, ..., y_n^{\text{o}})$ consists of positive-valued observed responses with the remaining k responses censored from below at 0. Assuming the GAL distribution for the latent responses, we can express the likelihood as $\prod_{i=1}^{n} f_{p_0}(y_i^{\mathrm{o}} \mid \gamma, \boldsymbol{x}_i^T\boldsymbol{\beta}, \sigma) \prod_{j=1}^{k} \int_{-\infty}^{0} f_{p_0}(w \mid \gamma, \boldsymbol{x}_{n+j}^T\boldsymbol{\beta}, \sigma)\,\mathrm{d}w$. Using data augmentation (Chib, 1992), we let $\boldsymbol{w} = (w_1, ..., w_k)$ be the unobserved (latent) responses corresponding to the k data points that are left-censored at 0. Then, under the hierarchical representation of the GAL distribution, the joint posterior distribution that includes $\boldsymbol{w}$ can be expressed as

$$p(\boldsymbol{\beta}, \gamma, \sigma, \{s_i\}, \{v_i\}, \boldsymbol{w} \mid \text{data}) \propto \pi(\boldsymbol{\beta}, \gamma, \sigma) \prod_{i=1}^{n} \mathrm{N}(y_i^{\mathrm{o}} \mid \boldsymbol{x}_i^T\boldsymbol{\beta} + \sigma C|\gamma| s_i + A v_i, \sigma B v_i)$$

$$\prod_{j=1}^{k} \mathrm{N}^{-}(w_j \mid \boldsymbol{x}_{n+j}^T\boldsymbol{\beta} + \sigma C|\gamma| s_{n+j} + A v_{n+j}, \sigma B v_{n+j}) \prod_{i=1}^{n+k} \mathrm{Exp}(v_i \mid \sigma^{-1})\, \mathrm{N}^{+}(s_i \mid 0, 1)$$

where $\pi(\boldsymbol{\beta}, \gamma, \sigma)$ denotes the prior for the model parameters, and $v_i = \sigma z_i$. Here, N^{-} denotes a truncated normal on $\mathbb{R}^{-}$, and $\mathrm{Exp}(v \mid \sigma^{-1})$ an exponential distribution with mean σ.

Regarding posterior inference, the posterior full conditional for each auxiliary variable w_j is given by a truncated normal distribution. Given the augmented data $(\boldsymbol{y}^{\mathrm{o}}, \boldsymbol{w})$, the model parameters and the latent variables $\{(v_i, s_i) : i = 1, ..., n+k\}$ can be sampled as in Section 2.2.1. We tested the posterior sampling algorithm on simulated data sets with sample size $n = 400$, generated from a simple regression setting with GAL errors with a censoring rate ranging from 20% to 40%. Under this scenario, the posterior distributions successfully captured the true values of all parameters in their 95% credible intervals.

2.3 Simulation study for regularized regression

Here, we present results from a simulation study designed to compare the lasso regularized quantile regression models with the AL and the GAL errors. We follow a standard simulation setting from the literature regarding the linear regression component (Tibshirani, 1996; Zou & Yuan, 2008; Li et al., 2010), varying the extent of sparsity in the true $\boldsymbol{\beta}$ vector. For the underlying data-generating error distributions, we consider four scenarios with different types of skewness and tail behavior. For model comparison, we evaluate the accuracy in variable selection based on the posterior inference under Bayesian Lasso shrinkage, the inference for the regression function, and the posterior predictive performance, using relevant assessment criteria. Overall, the GAL-based quantile regression model performs better in variable selection and prediction accuracy and it is more robust to non-standard error distributions, particularly for extreme quantiles. The two models yield comparable results in the case of median regression.

2.3.1 Simulation settings

We consider synthetic data generated from linear quantile regression settings, with $p_0 = 0.05$, 0.25 and 0.5 to study model performance for both extreme and more central percentiles. The rows of the design matrix were generated independently from an 8-dimensional normal distribution with zero mean vector and covariance matrix with elements $0.5^{|i-j|}$, for $1 \leq i, j \leq 8$. We present detailed results from a relatively sparse case for the vector of regression coefficients, $\boldsymbol{\beta} = (3, 1.5, 0, 0, 2, 0, 0, 0)$. In Section 2.3.3, we briefly discuss results form two other scenarios for $\boldsymbol{\beta}$ corresponding to a dense and a very sparse case.

Data were simulated under four different error distributions:

- $\mathrm{N}(\mu, 9)$, with μ chosen such that the p_0th quantile is 0.
- $\mathrm{Laplace}(\mu, 3)$, with μ chosen such that the p_0th quantile is 0.
- $0.1\mathrm{N}(\mu, 1) + 0.9\mathrm{N}(\mu + 1, 5)$, with μ chosen such that the p_0th quantile is 0.
- Log-transformed generalized $\mathrm{Pareto}(\sigma, \xi)$, with $\xi = 3$ and σ chosen such that the

p_0th quantile is 0. To generate the errors, we first sample from a generalized Pareto distribution, then take the logarithm. Based on the parameterization in Embrechts et al. (1997), the density function of the errors is given by $f(\epsilon \,|\, \sigma, \xi) = \sigma^{-1}\{1 + \xi\sigma^{-1}\exp(\epsilon)\}^{-(1+\xi^{-1})}\exp(\epsilon)$, for $\epsilon \in \mathbb{R}$.

The normal and Laplace error distributions are symmetric about zero under median regression. The parameters of the two-component normal mixture are selected such that the resulting error distribution is skewed. Finally, the log-transformed generalized Pareto distribution is included to study model performance under an error density which is both skewed and does not have exponential tails.

For each setting of the simulation study, we generated 100 data sets, each with $n = 100$ observations for training the models and another $N = 100$ for testing predictions.

2.3.2 Criteria for comparison

We consider a number of criteria to assess different aspects of model performance. Since Bayesian lasso regression only shrinks the covariate effects, we consider a threshold on the effect size for the purpose of variable selection. Following Hoti & Sillanpää (2006), we calculate the standardized effects as $\beta_j^* = (s_{x_j}/s_y)\beta_j$, $j = 1, \ldots, d$, where s_{x_j} is the standard deviation of predictor x_j and s_y is the standard deviation of the response. For each posterior sample, if the standardized effect is greater than 0.1 in absolute value, we consider the predictor as included. We count the number of correct inclusion and exclusions (CIE) in the posterior sample and divide it by d to normalize it to a number between 0 and 1. By averaging over all the posterior samples, we obtain the mean standardized CIE for each simulated data set.

To assess predictive performance for the regression function, we calculate the mean absolute deviation on $N = 100$ test data points, defined as: MAD $= \frac{1}{100}\sum_{i=1}^{100} |(\beta_0^* + \boldsymbol{x}_i^T\boldsymbol{\beta}^*) - \boldsymbol{x}_i^T\boldsymbol{\beta}|$, where β_0^* and $\boldsymbol{\beta}^*$ are the posterior mean estimate of the intercept and the regression vector from the training data. The MAD measures the average L_1 distance between the predicted quantile and the true quantile for the test data, thus can be viewed as a numeric evaluation of the posterior predictive inference under each model.

Finally to assess model fitting taking into account predictive uncertainty, we apply the posterior predictive loss criterion from Gelfand & Ghosh (1998). This criterion favors the model $\mathcal{M}$ that minimizes $D_m(\mathcal{M}) = P(\mathcal{M}) + \{m/(m+1)\}G(\mathcal{M})$, where $G(\mathcal{M}) = \sum_{i=1}^{n}\{y_i - \mathrm{E}^{\mathcal{M}}(y_i^* \mid \text{data})\}^2$ is a goodness-of-fit term, and $P(\mathcal{M}) = \sum_{i=1}^{n} \mathrm{var}^{\mathcal{M}}(y_i^* \mid \text{data})$ is a penalty term for model complexity. Here, $m \geq 0$, and $\mathrm{E}^{\mathcal{M}}(y_i^* \mid \text{data})$ and $\mathrm{var}^{\mathcal{M}}(y_i^* \mid \text{data})$ are the mean and variance under model $\mathcal{M}$ of the posterior predictive distribution for replicated response y_i^* with corresponding covariate $\boldsymbol{x}_i$. We also consider the generalized version of the criterion based on the check loss function, under which $D(\mathcal{M}) = \sum_{i=1}^{n} \mathrm{E}^{\mathcal{M}}(\rho_{p_0}(y_i - y_i^*) \mid \text{data})$. For this generalized criterion, the goodness-of-fit term can be defined by $G(\mathcal{M}) = \sum_{i=1}^{n} \rho_{p_0}(y_i - \mathrm{E}^{\mathcal{M}}(y_i^* \mid \text{data}))$ and the penalty term by $P(\mathcal{M}) = D(\mathcal{M}) - G(\mathcal{M})$, since the check loss function $L(y, a) \equiv \rho_{p_0}(y - a) = (y - a)p_0 - (y - a)I(y < a)$ is convex in y, and thus $P(\mathcal{M}) \geq 0$; see Gelfand & Ghosh (1998) for details on defining the model comparison criterion under loss functions different from quadratic loss.

2.3.3 Results

We use the same hierarchical Laplace prior for $\boldsymbol{\beta}$ under the AL and GAL models, with a gamma prior for η^2 with prior mean 1 and variance 10. Such prior specification is relatively non-informative in the sense that it does not favor shrinkage for the regression coefficients, resulting in marginal prior densities for each β_k that place substantial probability mass away from 0. The shape parameter γ of the GAL error distribution was assigned a uniform prior. Results under both models and for each simulated data set are based on 5,000 posterior samples, obtained after discarding the first 50,000 iterations of the Markov chain Monte Carlo sampler and then retaining one every 20 iterations.

Within each simulation scenario, we summarize results from the 100 data sets using the median and standard deviation (SD) of the values for the performance assessment criteria discussed in Section 2.3.2. Results are reported in Table 2.1 through Table 2.4, where we use boldface to indicate the model supported by the particular criterion under each setting.

Overall, the lasso regularized Bayesian quantile regression model performs better under the GAL error distribution. The GAL-based model includes/excludes correct regression coefficient values more often than the AL model for almost all combinations of p_0 and error distributions (Table 2.1). It also results in a lower median mean absolute deviation for the test data in most cases, demonstrating better performance in the prediction of the regression function (Table 2.2). Note that, for both types of assessment in Tables 2.1 and 2.2, the GAL-based model produces better results across all error distributions for $p_0 = 0.05$, and, with the exception of one case, when $p_0 = 0.25$. Results are generally more balanced

		Error distribution			
p_0	Model	Normal	Laplace	Normal mixture	log-transformed generalized Pareto
0.05	GAL	**0.848** (0.063)	**0.633** (0.083)	**0.911** (0.042)	**0.893** (0.052)
	AL	0.746 (0.099)	0.534 (0.087)	0.817 (0.075)	0.840 (0.081)
0.25	GAL	**0.851** (0.049)	**0.728** (0.060)	**0.918** (0.048)	0.896 (0.050)
	AL	0.843 (0.069)	0.700 (0.068)	0.913 (0.060)	**0.900** (0.051)
0.50	GAL	0.848 (0.052)	**0.738** (0.065)	**0.909** (0.049)	**0.897** (0.055)
	AL	**0.850** (0.056)	0.737 (0.065)	0.905 (0.050)	0.870 (0.061)

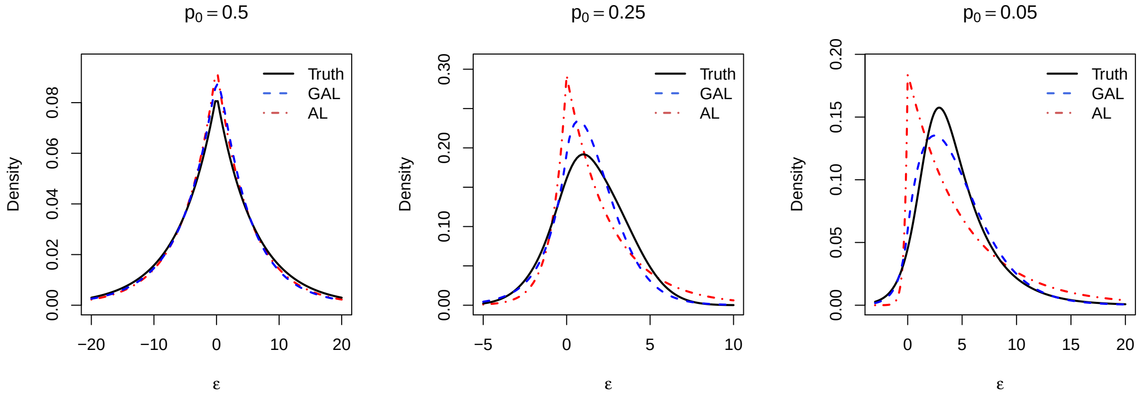
Table 2.1: Simulation study. Standardized number of correctly included/excluded predictors: median (SD).

		Error distribution			
p_0	Model	Normal	Laplace	Normal mixture	Log-transformed generalized Pareto
0.05	GAL	**0.899** (0.281)	**3.095** (0.831)	**0.596** (0.198)	**0.625** (0.173)
	AL	1.130 (0.280)	3.826 (1.012)	0.847 (0.218)	0.855 (0.254)
0.25	GAL	**0.705** (0.190)	**1.517** (0.470)	**0.503** (0.150)	**0.568** (0.165)
	AL	0.805 (0.191)	1.777 (0.506)	0.550 (0.159)	0.582 (0.172)
0.50	GAL	0.729 (0.185)	1.400 (0.415)	**0.512** (0.141)	**0.570** (0.155)
	AL	**0.710** (0.183)	**1.391** (0.416)	0.520 (0.132)	0.649 (0.183)

Table 2.2: Simulation study. MAD of regression function based on the test data: median (SD).

in the median regression setting, although the GAL model fares better in all cases for which the underlying error distribution is skewed.

For each simulation setting, Table 2.3 includes the values for the posterior predictive loss criterion with quadratic loss (under $m \to \infty$, such that $D_\infty = P + G$), and Table 2.4 shows the generalized criterion under check loss. Both versions of the posterior predictive loss criterion support the GAL model when $p_0 = 0.05$, with differences in values between the two models that are substantially larger than for the other two values of p_0. This reinforces the earlier findings on the potential benefits of the GAL error distribution for extreme percentiles. With the exception of one case under the check loss version of the



(a) Laplace error

(b) Mixture of two normals

(c) Log-generalized Pareto

Figure 2.2: Simulation example. Posterior predictive error density produced by fitting single simulated data sets with quantile lasso regularized model under AL and GAL errors. The data sets are simulated with different error distributions. From left to right the true error distributions have its 50th, 25th and 5th quantile equal to zero.

criterion, the GAL-based model is also favored when $p_0 = 0.25$, whereas results are more mixed in the median regression case.

We plot the posterior predictive error density to illustrate a visual comparison of the posterior inference under each model (Figure 2.2). Each panel shows the prediction from fitting a single simulated data set with both the AL and the GAL regression model. The data sets are generated under the Laplace distribution, mixture of normal distributions and log-transformed generalized Pareto distribution, with the 50th, 25th and 5th quantile set at zero, respectively. In all three examples, the model assuming GAL errors produces posterior predictive error density that is closer to the true error density than its AL counterpart.

We also considered two more settings for $\boldsymbol{\beta}$, a dense case with all 8 regression coefficients equal to 0.85, and a very sparse case with $\boldsymbol{\beta} = (5, 0, 0, 0, 0, 0, 0, 0)$. The conclusions were overall similar, in particular, the GAL model outperformed the AL model for essentially all combinations of underlying error distribution and value of $p_0 = 0.05$ or

			Error distribution			
p_0	Model	Score	Normal	Laplace	Normal mixture	log-transformed generalized Pareto
0.05	GAL	P	**1231** (193)	**9799** (2483)	**653** (112)	**1273** (270)
		G	**832** (126)	**7046** (1546)	**429** (71)	**1053** (267)
		D_∞	**2092** (312)	**16860** (3839)	**1085** (181)	**2319** (531)
	AL	P	3359 (799)	30308 (10763)	1839 (405)	2782 (664)
		G	952 (165)	8659 (2304)	534 (93)	1168 (279)
		D_∞	4357 (933)	38766 (12676)	2398 (487)	4020 (873)
0.25	GAL	P	**1085** (206)	**6977** (1607)	**608** (95)	**1445** (273)
		G	**830** (146)	**6897** (1606)	**444** (66)	**1105** (264)
		D_∞	**1882** (343)	**13884** (3115)	**1055** (154)	**2552** (511)
	AL	P	1630 (303)	11503 (2727)	884 (148)	1516 (260)
		G	865 (154)	7395 (1742)	464 (71)	1113 (263)
		D_∞	2499 (448)	18916 (4349)	1352 (215)	2600 (487)
0.50	GAL	P	1283 (205)	7600 (1676)	694 (97)	**1189** (217)
		G	**813** (132)	6459 (1509)	**424** (60)	**1089** (245)
		D_∞	2111 (328)	14076 (3101)	1121 (152)	**2283** (415)
	AL	P	**1177** (191)	**7256** (1572)	**634** (87)	1318 (247)
		G	818 (134)	**6431** (1509)	426 (60)	1107 (255)
		D_∞	**2008** (318)	**13667** (3019)	**1058** (143)	2415 (483)

Table 2.3: Simulation study. Penalty term (P), goodness-of-fit term (G) and posterior predictive loss criterion (D_∞) under quadratic loss: median (SD).

$p_0 = 0.25$. Again, in the median regression case, the distinction between the two models was less clear for the normal, Laplace and normal mixture data-generating distributions, although the GAL model performed better under all criteria for the setting corresponding to the log-transformed generalized Pareto distribution.

2.4 Data examples

In this section, we consider three data examples to illustrate the Bayesian quantile regression models developed in Sections 2.2.1, 2.2.2, and 2.2.3. The main emphasis is on the comparison of inference results between models based on the GAL distribution and those

		Error distribution			
p_0	Model	Normal	Laplace	Normal mixture	log-transformed generalized Pareto
0.05	GAL	**174.2** (13.6)	**507.0** (67.8)	**122.8** (11.3)	**178.5** (17.4)
	AL	209.3 (21.7)	605.4 (70.8)	148.6 (17.3)	200.2 (20.5)
0.25	GAL	**169.5** (15.9)	**443.9** (47.1)	**126.2** (9.5)	188.0 (17.5)
	AL	178.0 (15.7)	451.4 (45.8)	129.0 (9.8)	**185.5** (17.3)
0.50	GAL	175.7 (13.4)	444.7 (48.0)	127.4 (8.9)	**178.6** (16.1)
	AL	**172.6** (13.4)	**438.5** (47.5)	**125.2** (8.7)	183.6 (18.1)

Table 2.4: Simulation study. Posterior predictive loss criterion under check loss: median (SD).

assuming an AL distribution for the errors.

We have implemented both models with priors for their parameters that result in essentially the same prior predictive error densities. The two models were applied with the same prior distributions for $\boldsymbol{\beta}$ and σ. For the data sets of Sections 2.4.1 and 2.4.3, we used a $\mathrm{N}(\mathbf{0}, 100I)$ prior for the vector of regression coefficients, and an $\mathrm{IG}(2, 2)$ prior for the scale parameter σ. For the data example of Section 2.4.2, we used a $\mathrm{N}(0, 100)$ prior for the intercept, and the same conditional Laplace prior for the remaining regression coefficients with the simulation study (see Section 2.3.3). Finally, a uniform prior was placed on the shape parameter γ of the GAL error distribution. For all data examples, the posterior densities for model parameters were fairly concentrated relative to the corresponding prior densities.

2.4.1 Immunoglobulin-G data

We illustrate the proposed model, referred to as model M_1, with a data set commonly used in additive quantile regression; see, for instance, Yu & Moyeed (2001). The analysis focuses on comparison with the simpler model based on asymmetric Laplace er-

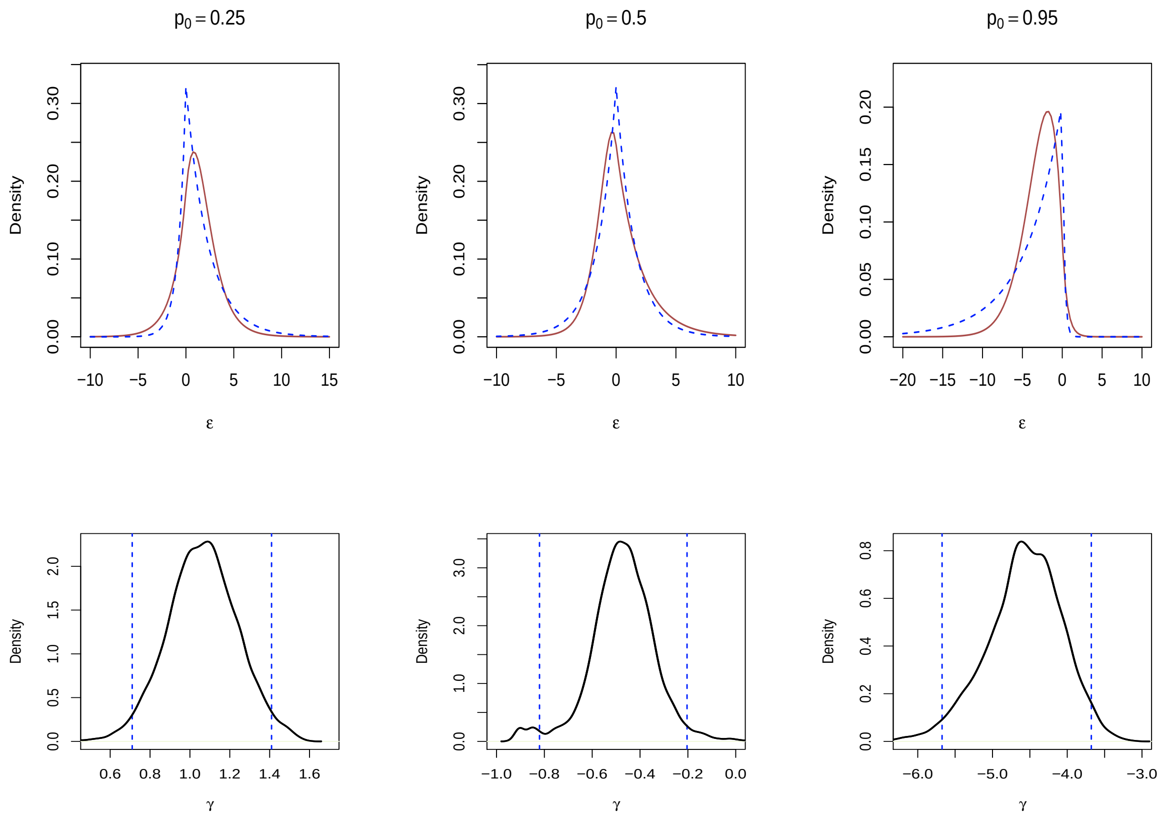


Figure 2.3: Immunoglobulin-G data. Inference results for $p_0 = 0.25$, 0.5 and 0.95. Top row: posterior predictive error densities under the asymmetric Laplace model (dashed lines) and the generalized asymmetric Laplace model (solid lines). Bottom row: posterior densities for parameter γ, with the vertical lines corresponding to the endpoints of the 95% credible interval.

rors, referred to as model M_0. The data set contains the immunoglobulin-G concentration in grams per litre for $n = 298$ children aged between 6 months and 6 years. As in earlier applications of quantile regression for these data, we use a quadratic regression function $\beta_0 + \beta_1 x + \beta_2 x^2$ to model five quantiles, corresponding to $p_0 = 0.05$, 0.25, 0.5, 0.75, 0.95, of immunoglobulin-G concentration against covariate age (x).

The two models result in different posterior predictive error densities, especially for extreme percentiles; see Figure 2.3. At $p_0 = 0.95$, under the AL model, both the shape

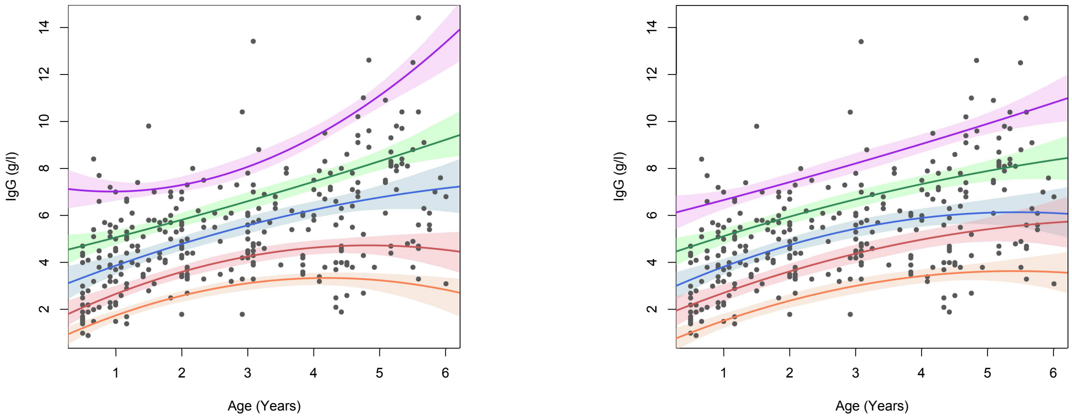


Figure 2.4: Immunoglobulin-G data. Posterior mean estimates and 95% credible bands for the quantile regression function $\beta_0 + \beta_1 x + \beta_2 x^2$ against age (x), for $p_0 = 0.05$, 0.25, 0.50, 0.75 and 0.95. Left: AL model. Right: GAL model.

and the skewness of the error distribution are predetermined by p_0 and the mode is forced to be 0, resulting in a rigid heavy left tail. The effect of this overly dispersed tail can be observed in the inference for the quantile regression function (Figure 2.4). The GAL model, on the contrary, yields an error density that has a much thinner left tail, concentrating more of its probability mass around the mode, which is not at 0. Figure 2.3 also shows the posterior densities for shape parameter γ, under a uniform prior in all cases. For all three quantile regressions, the 95% posterior credible interval for γ does not include the value of 0, which corresponds to asymmetric Laplace errors. Median regression is the only case where 0 is within the effective range of the posterior distribution for γ.

For formal model comparison, we compute the Bayesian information criterion (BIC), the posterior predictive loss criterion with quadratic loss, and the generalized posterior predictive loss criterion under the check loss. We further calculate the approximate

		Approximate Bayes Factor		Bayesian information criterion	
Quantile	Model	log-BF$_{10}$	BF$_{10}$	log-likelihood	BIC
$p_0 = 0.05$	M$_0$	–	–	−666	1355
	M$_1$	60.9	2.9×10^{26}	−615	1258
$p_0 = 0.25$	M$_0$	–	–	−632	1287
	M$_1$	11.7	1.2×10^{5}	−622	1273
$p_0 = 0.50$	M$_0$	–	–	−633	1289
	M$_1$	8.9	7.4×10^{3}	−623	1274
$p_0 = 0.75$	M$_0$	–	–	−654	1331
	M$_1$	37.8	2.6×10^{16}	−620	1268
$p_0 = 0.95$	M$_0$	–	–	−761	1545
	M$_1$	127.4	2.2×10^{55}	−646	1320

Table 2.5: Immunoglobulin-G data. Approximate Bayes factor and Bayesian information criterion under the asymmetric Laplace and generalized asymmetric Laplace models, denoted by M$_0$ and M$_1$, respectively.

Bayes factor of the model under the proposed distribution (M$_1$) versus that under the asymmetric Laplace distribution (M$_0$) using Laplace approximation (Raftery, 1996), denoted as $B_{10} = P(D \mid M_1)/P(D \mid M_0)$ with D as the data. Here, $P(D \mid M_k) = \int P(D \mid \boldsymbol{\theta}_k, M_k)P(\boldsymbol{\theta}_k \mid M_k)\,\mathrm{d}\boldsymbol{\theta}_k$, where $\boldsymbol{\theta}_k$ is the parameter set of M_k. Under Laplace approximation, $P(D \mid M_k) \approx (2\pi)^{d_k/2}|\hat{H}_k|^{-1/2}P(D \mid \hat{\boldsymbol{\theta}}_{\boldsymbol{k}}, M_k)P(\hat{\boldsymbol{\theta}}_{\boldsymbol{k}} \mid M_k)$, where d_k and $\hat{\boldsymbol{\theta}}_{\boldsymbol{k}}$ are the dimension and the posterior mode of $\boldsymbol{\theta}_k$, $\hat{H}_k$ being the Hessian of $\log\{P(D \mid \boldsymbol{\theta}_k, M_k)P(\boldsymbol{\theta}_k \mid M_k)\}$ evaluated at $\hat{\boldsymbol{\theta}}_{\boldsymbol{k}}$ (Raftery, 1996). We approach $P(D \mid M_k)$ with approximations because both integrals involve high dimensions and are challenging to compute directly. More specifically, we calculate the denominator (M$_0$) with Laplace approximation and obtain the numerator (M$_1$) with a Riemann sum of the Laplace approximation of the integral given γ over an evenly-spaced grid of γ.

Both the BIC and the Bayes factor favor the new model at all five quantiles; see Table 2.5. Under the posterior predictive loss criterion (Table 2.6), the two models are comparable in the case of median regression, with model M_0 preferred. In all other cases, model M_1 is favored by both versions of the model comparison criterion. The improvement in performance over the AL model is particularly conspicuous at the two extreme percentiles. This is in agreement with the difference in the posterior predictive error densities for $p_0 = 0.95$, reported in Figure 2.3.

		Posterior predictive loss criterion					
		Quadratic loss			Check loss		
Quantile	Model	P	G	D_∞	P	G	D
$p_0 = 0.05$	M_0	3511	1331	4841	179	180	359
	M_1	1298	1170	2467	230	102	331
$p_0 = 0.25$	M_0	1820	1180	3001	232	123	355
	M_1	1407	1144	2551	236	108	343
$p_0 = 0.50$	M_0	1465	1142	2607	229	108	338
	M_1	1626	1161	2788	232	114	346
$p_0 = 0.75$	M_0	2122	1227	3350	201	134	335
	M_1	1208	1140	2348	228	97	325
$p_0 = 0.95$	M_0	6522	1751	8273	137	259	395
	M_1	1525	1165	2690	208	118	327

Table 2.6: Immunoglobulin-G data. Posterior predictive loss criterion (based on quadratic loss and check loss functions) under the asymmetric Laplace and generalized asymmetric Laplace models, denoted by M_0 and M_1.

2.4.2 Boston housing data

We apply the lasso regularized quantile regression model to the realty price data from the Boston Standard Metropolitan Statistical Area (SMSA) in 1970 (Harrison & Rubinfeld, 1978). The data set contains 506 observations. We take the log-transformed corrected median value of owner-occupied housing in USD 1000 (LCMEDV) as the response, and consider the following predictors: point longitudes in decimal degrees (LON), point latitudes in decimal degrees (LAT), per capita crime (CRIM), proportions of residential

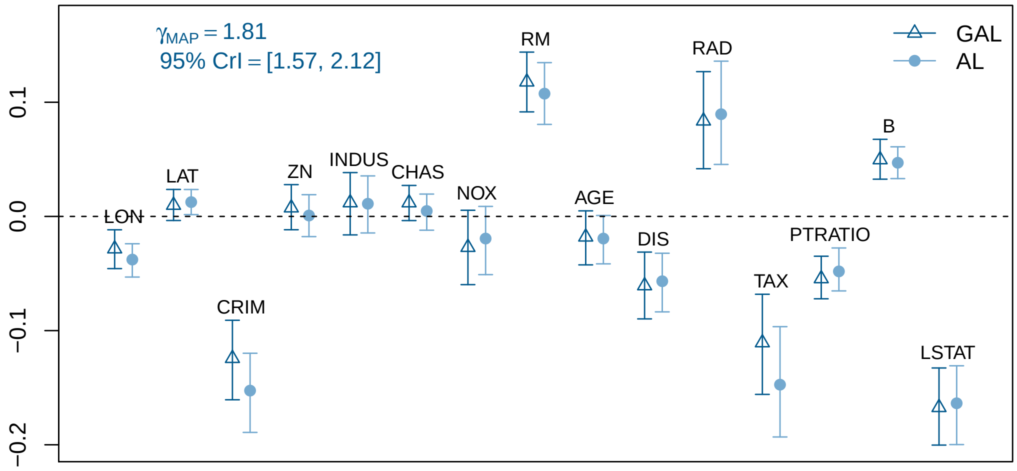


Figure 2.5: Boston housing data. Posterior point and 95% interval estimates for the regression coefficients of the 10th quantile lasso regularized model under AL and GAL errors.

land zoned for lots over 25000 square feet per town (ZN), proportions of non-retail business acres per town (INDUS), a factor indicating whether tract borders Charles River (CHAS), nitric oxides concentration (parts per 10 million) per town (NOX), average numbers of rooms per dwelling (RM), proportions of owner-occupied units built prior to 1940 (AGE), weighted distances to five Boston employment centers (DIS), index of accessibility to radial highways per town (RAD), full-value property-tax rate per USD 10,000 per town (TAX), pupil-teacher ratios per town (PTRATIO), transformed African American population proportion (B), and percentage values of lower status population (LSTAT).

We consider quantiles of 0.1 and 0.9 and compare the maximum a posteriori estimates (MAP) of regression coefficients, along with 95% credible intervals, for standardized covariates under the lasso regularized quantile regression models with AL and GAL errors (Figure 2.5 and 2.6). For both quantiles, the widths of the 95% credible intervals for the regression coefficients are overall comparable between the two models, but the posterior point estimates can be quite different. For instance, under the 10th quantile regression, the

GAL model shrinks the effects of per capita crime (CRIM) and property-tax rate (TAX) to a greater extent compared to the AL model. Similar patterns can be observed for index of accessibility to radial highways (RAD) for the 90th quantile. Moreover, the two models reach different conclusions on the effect of latitude (LAT) for the 10th percentile. Although the posterior point estimates suggest a higher housing price as latitude increases adjusting for all other covariates, the 95% credible interval under the GAL model includes 0, whereas the one under the AL model does not.

Focusing on inference under the GAL error distribution, we note that, although the model selected some common variables for the two quantiles, there is also some discrepancy. For instance, each of higher proportions of residential land zoned for lots over 25000 square feet per town (ZN) and having tracts bordering Charles river (CHAS) increase the price at the 90% percentile, while higher nitrogen oxide value (NOX) has a negative influence on the 90% percentile price. However, none of these covariates have a significant effect on the realty value at the 10% percentile.

Finally, we notice that for both the 10th and 90th quantile regression, 0 is far away from the endpoints of the 95% credible interval for the GAL model shape parameter γ. This suggests that asymmetric Laplace errors are not suitable for this particular application. This is further supported by the results for the posterior predictive loss criterion reported in Table 2.7.

2.4.3 Labor supply data

We illustrate the Tobit quantile regression model with the female labor supply data from Mroz (1987), which was taken from the University of Michigan Panel Study of

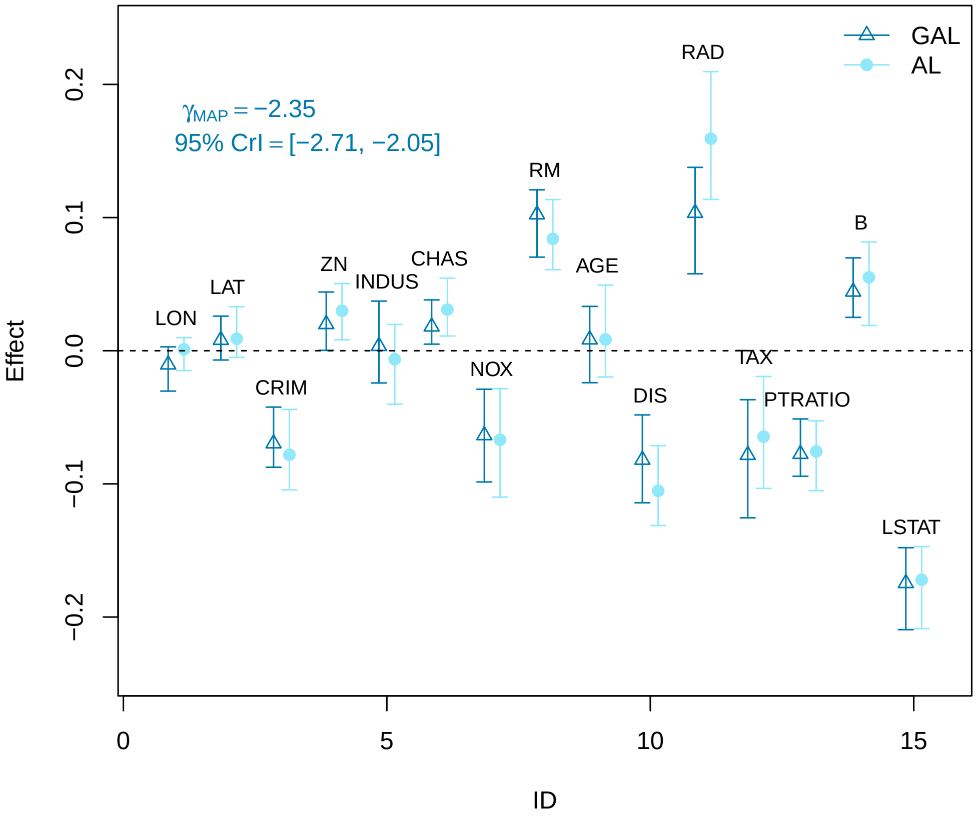


Figure 2.6: Boston housing data. Posterior point and 95% interval estimates for the regression coefficients of the 90th quantile lasso regularized model under AL and GAL errors.

		Posterior predictive loss criterion					
		Quadratic loss			Check loss		
Quantile	Model	P	G	D_∞	P	G	D
$p_0 = 0.10$	M_0	46.9	26.2	73.1	28.1	22.6	50.7
	M_1	22.8	20.1	42.9	30.4	18.5	48.9
$p_0 = 0.90$	M_0	74.8	28.8	103.6	24.1	31.5	55.7
	M_1	22.6	18.4	41.0	26.3	21.0	47.3

Table 2.7: Boston housing data. Posterior predictive loss criterion (based on quadratic loss and check loss functions) under the AL (model M_0) and GAL (model M_1) error distribution.

Income Dynamics for year 1975. The data set includes records on the work hours and other relevant information of 753 married white women aged between 30 and 60 years old. Of the 753 women, 428 worked at some time during 1975, with the corresponding fully observed responses given by the wife's work hours (in 100 hours). For the remaining 325 women, the observed zero work hours correspond to negative values for the latent "labor supply" response. We use the quantile regression function considered in Kozumi & Kobayashi (2011), where an AL-based Tobit quantile regression model was applied to the same data set. The linear predictor includes an intercept, income which is not due to the wife (`nwifeinc`), education of the wife in years (`educ`), actual labor market experience in years (`exper`) and its quadratic term (`expersq`), age of the wife (`age`), number of children less than 6 years old in household (`kidslt6`), and number of children between ages 6 and 18 in household (`kidsge6`). We compare the results from the Bayesian Tobit quantile regression model assuming AL errors (model M_0) and GAL errors (model M_1).

Table 2.8 summarizes the posterior distribution of γ under the GAL model, and presents results from criterion-based comparison of the two models for $p_0 = 0.05$, 0.50 and 0.95. Since there is censoring in the data, we use the revised BIC from Volinsky & Raftery (2000). In all three cases, the 95% credible interval for γ excludes 0, and the GAL-based model is associated with lower BIC values. The results support the GAL-based model more emphatically for the extreme percentiles than for median regression.

Figure 2.7 shows the posterior distributions of labor supply quantiles corresponding to $p_0 = 0.05$, 0.50 and 0.95 for women with 0, 1, 2 and 3 children less than 6 years old. For all other predictors, we use the median values from the data as input values to represent

Quantile	Model	Mean (95% CrI) for γ	likelihood	BIC
$p_0 = 0.05$	M_0		-1975	4004
	M_1	5.22 (4.43, 6.24)	-1874	3809
$p_0 = 0.50$	M_0		-1867	3789
	M_1	0.58 (0.39, 0.81)	-1845	3750
$p_0 = 0.95$	M_0		-1967	3989
	M_1	-4.16 (-5.5, -3.06)	-1854	3769

Table 2.8: Labor supply data. Posterior mean and 95% credible interval for the shape parameter γ of the GAL error distribution, and BIC values under the AL and GAL models, denoted by M_0 and M_1, respectively.

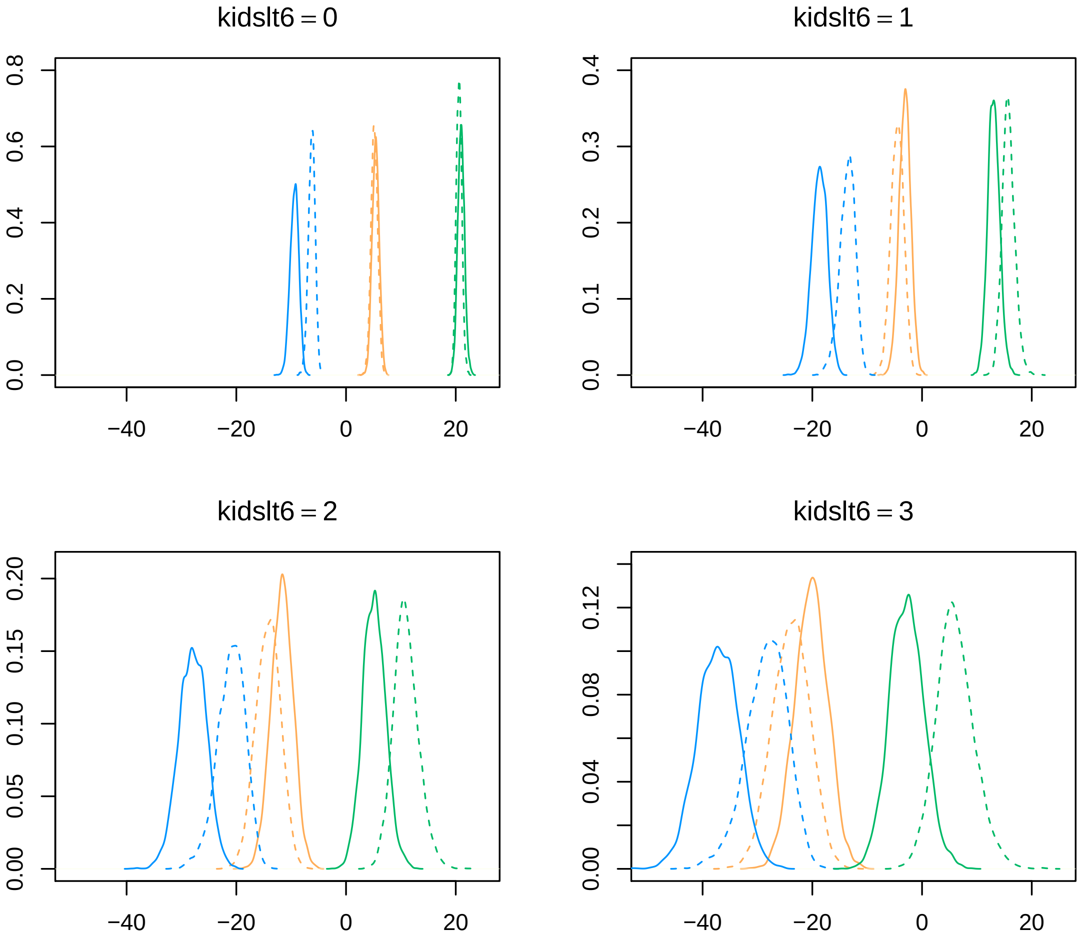


Figure 2.7: Labor supply data. Posterior densities for the 5th (blue), 50th (orange) and 95th quantile (green) of labor supply (in 100 hours) for women with 0, 1, 2 or 3 children less than 6 years old. The solid (dashed) lines correspond to the posterior densities under the GAL (AL) model.

an *average* wife. As the number of young children increases, the AL model estimates the 5th quantile and the median of labor supply of an average wife to be closer to each other. Under the GAL model, the distance between the densities of the 5th quantile and median labor supply also decreases with increasing number of young children, albeit at a lower rate. When estimating the 95th quantile, the proposed model is more conservative than the AL model about the labor contribution of an average wife with an increasing number of children less than 6 years old. When there are 3 children less than 6 years old in the household, the center of the posterior distribution for the 95th quantile is below zero under the GAL model, meaning that even at the top 5th percentile of labor supply, an average wife may still produce negative labor supply as she takes care of many young family members. More specifically, the posterior probability of the 95th labor supply quantile being positive is 0.19 under the GAL model, as opposed to 0.97 under the AL model. These results demonstrate that the choice of error distribution in quantile regression can have an effect on practically important conclusions for a particular application.

2.5 Discussion

We have developed a Bayesian quantile regression framework with a new error distribution that has flexible skewness, mode and tail behavior. The proposed model has better performance compared with the commonly used asymmetric Laplace distribution, particularly for modeling extreme quantiles. Owing to the hierarchical structure of the new distribution, posterior inference and prediction can be readily implemented via Markov chain Monte Carlo methods.

The main motivation for the work in this chapter was to develop a sufficiently flexible parametric distribution that can be used as a building block for different types of quantile regression models. The extension to quantile regression with ordinal responses is a possible direction. Expanding the model to a spatial quantile regression process, along the lines of Lum et al. (2012), is another direction. More importantly for our objectives, this work lays the foundation for developing a Bayesian framework for mixtures of quantile regression components, which is the topic of next chapter.

Chapter 3

Bayesian quantile mixture regression

In this chapter, we propose a new regression procedure named Bayesian quantile mixture regression (BQMR) constructed with a weighted mixture of K regression components. By construction, each component is a p_kth quantile regression, with $k = 1, \ldots, K$ and $0 < p_1 < \ldots < p_K < 1$; and all components share a common regression coefficient vector $\boldsymbol{\beta}$. We present the idea of BQMR in Section 3.1 and discuss the objectives of two versions of the framework. The first version consists of mixtures of GAL_{p_k} distributions with known p_k, while the second is a mixture of AL_{p_k} components of which the percentage p_k are treated as random parameters.

The rest of this chapter is organized as follows. In Section 3.2, we formulate BQMR with mixtures of weighted GAL components for fixed $\{p_k\}$. The methodology covers the parametric framework and posterior inferences, as well as a semi-parametric extension with Dirichlet process prior on the weights. Results from simulation studies are presented in Section 3.3 to illustrate the performance of the models. In Section 3.4, we introduce the BQMR model with a mixture of AL distributions where p_k are random quantities. We further explore the scenario where the total number of components K is random and develop posterior inferences on K. A simulation study is conducted under this version of the framework in Section 3.5. Finally, we illustrate the proposed models using the Boston housing data example in Section 3.6 and conclude with a discussion in Section 3.7.

3.1 Idea and objectives

The idea of BQMR framework is to model the response variable with a finite weighted mixture of quantile regressions. Conceptually, the model can be expressed as

$$y_i \mid \boldsymbol{x}_i, \boldsymbol{\beta}, \boldsymbol{\theta}_1, \ldots, \boldsymbol{\theta}_K \quad \sim \quad \sum_{k=1}^{K} \omega_k Q_{p_k}(y_i \mid \boldsymbol{x}_i \boldsymbol{\beta}, \boldsymbol{\theta}_k)$$

where ω_k are weights of each regression component and Q_{p_k} is some density function such that the p_kth percentile of y_i is equal to the regression function $\boldsymbol{x}_i\boldsymbol{\beta}$, with $\boldsymbol{\theta}_k$ denoting all the remaining parameters of Q_{p_k}. We specify a common regression coefficient vector $\boldsymbol{\beta}$ for all K regressions to estimate the covariate effect synthesized from different quantile components. Through placing a sparsity-inducing prior on $\boldsymbol{\beta}$, we hope to achieve efficient shrinkage of the predictor effect by combining information from multiple quantile components, each depicting a different part of the response distribution.

A characterizing feature of this mixture framework is its kernel density Q. Given that the mixture model consists of quantile regression components, we need flexible kernel densities that are parameterized in terms of percentiles. One could potentially construct the mixture with AL kernel densities. However, this turns out to be a suboptimal choice if we want to fix $\{p_k\}$. Since the skewness of the AL distribution is fully determined by the percentage p_k of the quantile, the mixture model formed by AL distributions has very limited flexility under fixed $\{p_k\}$. Empirical demonstrations of this limitation of the AL kernel will be presented in the following sections.

We propose two approaches that overcome this issue. Firstly, with fixed $\{p_k\}$, we use the generalized asymmetric Laplace (GAL) distribution as the kernel density. As a generalization of the AL distribution, the GAL distribution has an extra parameter that allows for varying skewness and shape under given p_k. Alternatively, we formulate the mixture components with AL densities, but in this case we treat the percentage p_k of each quantile as random quantities. In other words, we only specify the total number of components K, but allow the values of p_k to be estimated from the data. By doing so, we relax the constraints on the skewness and shape of the AL components and attain improved flexibility of the mixture model. Under the AL framework, we will also consider the extension with random K and discuss posterior inference in this scenario.

The two models have different objectives. In the situation where we know which $\{p_k\}$ we are interested in, we can apply the GAL framework that allows us to fix the percentages at the desired values. In the absence of such information, the GAL framework under equally-spaced $\{p_k\}$ will give us an approximation of the response distribution given the covariates. The random-p_k AL framework applies to scenarios where we have minimal prior knowledge of which parts of the distribution may be more important for the prediction, but would like to obtain inferences on the $\{p_k\}$ of the components and to explore the response distribution in an efficient manner. In the simulation examples of this chapter, we will show that the two mixture models substantially outperform the simpler model configured with AL kernel densities and fixed p_k. A version of this model was considered in Huang & Chen (2015).

3.2 Bayesian quantile mixture regression with fixed p_k

In this section, we present the BQMR framework as a mixture of K weighted GAL distributions, each parameterized in terms of the p_kth quantile, where $K \geq 2$ is known and the percentages $\{p_k\}$ are specified and ordered, such that $0 < p_1 < \ldots < p_K < 1$. Details on the formulation and the mixture representation of the GAL distribution for the p_kth quantile can be found in Section 2.1 and Section 2.2. We elaborate on the choices of priors for the parameters, which includes a semi-parametric extension of the model with a Dirichlet process prior for the weights. Simulation examples are provided to facilitate the explanation of the construction and the inference under the framework.

3.2.1 Model formulation

We propose the BQMR with fixed p_k as a mixture of GAL_{p_k} distributions, where $\{p_k\}$ satisfy $0 < p_1 < p_2 < \ldots < p_K < 1$, to model the conditional distribution of response y. In a regression setting, the kth component is given by a p_kth-quantile GAL distribution with shape parameter γ_{p_k}, location $\mu_{p_k} + \boldsymbol{x}^T \boldsymbol{\beta}$ and scale parameter σ. By construction, all K components share the common regression coefficient $\boldsymbol{\beta}$ and scale σ. Denoting w_1, $\ldots$,

w_K as weights, we can write the model as

$$f(y \mid \boldsymbol{\beta}, \sigma, \{\omega_k\}, \{\gamma_{p_k}\}, \{\mu_{p_k}\}) \quad = \quad \sum_{k=1}^{K} \omega_k f_{p_k}^{GAL}(y \mid \gamma_{p_k}, \mu_{p_k} + \boldsymbol{x}^T\boldsymbol{\beta}, \sigma) \qquad (3.1)$$

where $\sum_{k=1}^{K} \omega_k = 1$ with $0 < \omega_k < 1$, $k = 1, \ldots, K$ and $f_{p_k}^{GAL}$ represents the density of the GAL_{p_k} distribution.

As the intercept in quantile regression, μ_{p_k} corresponds to the p_kth quantile of the kth component when all covariates are set to zero. To ensure ordering of the mixture components, we place a monotonicity constraint on $\{\mu_{p_k}\}$ by enforcing $\mu_{p_1} < \ldots < \mu_{p_k} < \ldots < \mu_{p_K}$, so that the p_kth quantile of the kth component is bounded from below by the p_{k-1}th quantile of the $(k-1)$th component and from above by the p_{k+1}th quantile of the $(k+1)$th component.

The motivation and the aim of this weighted mixture model contribute to the construction as well as the interpretation of the framework. In the case where we know a priori that particular parts of the response distribution, such as the center or the upper tail, provide the most information, we can select the corresponding quantile components for the mixture model. Then (3.1) can be viewed as a mixture model that captures the heterogeneity in the generation of the data. On the other hand, if we are lacking such prior information, we can construct the model with a good number of equally spaced p_k as an approximation of the true density. Moreover, by adopting the GAL_{p_k} distribution instead of the AL_{p_k} as kernel density, we improve the flexibility of the model under fixed p_k and enhance the robustness of the framework.

Owing to the mixture representation of the GAL distribution in (2.2), we can write the model in a hierarchical way by augmenting the parameter space with latent variables

z_i and s_i. For convenience of sampling, we set $v_i = \sigma z_i$ for $i = 1, \dots, n$ as in Kozumi & Kobayashi (2011) and introduce binary indicator variables ξ_{ik}, $i = 1, \dots, n$, $k = 1, \dots, K$ to break the additive mixture into products, where $\xi_{ik} = 1$ if the ith observation is allocated to the kth component and zero otherwise. Then the model can be expressed as:

$$\begin{aligned} y_i \mid \boldsymbol{\beta}, \boldsymbol{\mu}, \boldsymbol{\gamma}, \sigma, v_i, s_i, \boldsymbol{\xi}_i &\sim \prod_{i=1}^{K} \left[N(y_i \mid \mu_{p_k} + \boldsymbol{x}_i^T \boldsymbol{\beta} + \sigma C_{p_k} |\gamma_{p_k}| s_i + A_{p_k} v_i, \sigma B_{p_k} v_i) \right]^{\xi_{ik}} \\ \boldsymbol{\xi}_i \mid \omega_1, \dots \omega_k &\sim Multinomial(\boldsymbol{\xi}_i \mid \omega_1, \dots \omega_k) \\ v_i, s_i \mid \sigma &\overset{\text{i.i.d}}{\sim} Exp(v_i \mid 1/\sigma) N^+(s_i \mid 0, 1) \end{aligned}$$

where A_{p_k}, B_{p_k} and C_{p_k} are functions of shape γ_{p_k} and percentage p_k as are defined for (2.2) and γ_{p_k} takes value over (L_k, U_k), with $g(L_k) = 1 - p_0$, $L_k < 0$ and $g(U_k) = p_0$, $U_k > 0$.

3.2.2 Priors and posterior inference

The intercept μ_{p_k} and the shape parameter γ_{p_k} require component-specific priors. To respect the monotonicity constraint on the $\{\mu_{p_k}\}$, we construct the priors in a Markovian fashion, beginning with a $N(0, \sigma_\mu^2)$ prior for μ_{p_1} with some known σ_μ^2. Given μ_{p_1}, we sample $\mu_{p_2} | \mu_{p_1}$ from $N(0, \sigma_\mu^2) \mathbb{1}\{\mu_{p_2} > \mu_{p_1}\}$, which is a truncated normal distribution over (μ_{p_1}, ∞). The construction continues in this fashion for $2 \leq k \leq K$, so that given the previous intercept, μ_{p_k} follows $N(0, \sigma_\mu^2) \mathbb{1}\{\mu_{p_k} > \mu_{p_{k-1}}\}$. The ordering in μ_{p_k} is automatically satisfied through this sequential construction. We apply independent rescaled-Beta$(\alpha_\gamma, \beta_\gamma, L_k, U_k)$ prior for the shape parameters γ_{p_k} such that $\log(\frac{\gamma_{p_k} - L_k}{U_k - L_k})$ follows a Beta$(\alpha_\gamma, \beta_\gamma)$ distribution. When $\alpha_\gamma = \beta_\gamma = 1$, the prior is uniform over (L_k, U_k). On the other hand, if $\alpha_\gamma, \beta_\gamma > 1$, then the prior favors values that are not too close to either endpoints of the interval.

A natural prior for the common regression coefficients $\boldsymbol{\beta}$ is a Gaussian distribution.

Note that if instead we use a Laplace prior, the framework transforms into a tool for variable selection guided by the posterior intervals of covariate effects under shrinkage. In Section 2.2.2, we have shown that it is easy to achieve conditional Bayesian lasso under the GAL distribution by placing a Laplace prior with tuning parameter λ on the d-dimensional regression coefficient $\boldsymbol{\beta}$,

$$\pi(\boldsymbol{\beta} \mid \sigma, \lambda) = \prod_{j=1}^{d} \frac{\lambda}{2\sigma} \exp\left\{-\frac{\lambda}{\sigma}|\beta_j|\right\} = \prod_{j=1}^{d} \int_{\mathbb{R}^+} \frac{1}{\sqrt{2\pi\tau_j}} \exp\left\{-\frac{\beta_j^2}{2\tau_j}\right\} \frac{\eta^2}{2} \exp\left\{-\frac{\eta^2}{2}\tau_j\right\} \mathrm{d}\tau_j$$

The posterior sampling is based on the mixture representation of Laplace distribution by introducing the latent parameters τ_j and hyperparameter η^2. The latter can be learned from the data if complemented with a hyperprior.

Conjugate priors can be applied for the remaining parameters. The scale parameter σ receives an inverse-gamma prior, $\mathrm{IG}(a_\sigma, b_\sigma)$, with mean $b_\sigma/(a_\sigma - 1)$ provided $a_\sigma > 1$. We use a Dirichlet$(a_1, \ldots, a_K)$ prior for the weights $(\omega_1, \ldots, \omega_K)$, where equal a_k values imply equal weights on average a priori.

It is possible to make the model more flexible by placing a non-parametric prior on $\{\omega_k\}$. Motivated by the weight construction in the Bernstein prior (Petrone, 1999), we define $\omega_k = G(p_k) - G(p_{k-1})$ for $k = 2, \ldots, K$ and $\omega_1 = G(p_1)$, where G follows a Dirichlet process (DP) with concentration parameter α_0 and baseline distribution G_0 supported over $[0, p_K]$. Following the definition of Dirichlet process in Ferguson (1973), the prior of $\{\omega_k\}$ can be written as,

$$\omega_1, \omega_2, \ldots, \omega_K \mid \alpha_0, G_0 \quad \sim \quad Dir\left(\alpha_0 G_0(p_1), \alpha_0[G_0(p_k) - G_0(p_{k-1})], \ldots, \alpha_0[1 - G_0(p_{K-1})]\right)$$

Since the baseline distribution is defined over $[0, p_K]$, we consider a rescaled-Beta$(\alpha_G, \beta_G, 0, p_K)$ as G_0, such that x/p_K follows a Beta(α_G, β_G). We parameterize the

Beta distribution under its mean $\mu_G \in (0, 1)$ and the scale $\tau_G > 0$, so that $\alpha_G = \tau_G \mu_G$ and $\beta_G = \tau_G(1 - \mu_G)$. The scale parameter τ_G is fixed to some known constant, the choice of which conveys the prior knowledge of the spread of the baseline distribution. The larger τ_G is, the more concentrated G_0 will be in the case of $\mu_G = 0.5$. The mean parameter μ_G is random and will inform the center of the realizations of DP. We estimate μ_G by applying a Beta(α_μ, β_μ) prior with known α_μ and β_μ.

We implement MCMC sampling to fit the model and explore the posterior distribution. Under the parametric Dirichlet distribution prior for the weights, the MCMC algorithm consists of an adaptive Metropolis-Hastings step for γ_{p_k} and Gibbs steps for all the other parameters (details provided in Appendix A.2). Compared with the parametric BQMR, the posterior samples in the nonparametric extension can be obtained in a similar fashion except for a slightly different Gibbs update for $\{w_k\}$, plus an additional Metropolis-Hastings step to sample the mean parameter μ_G of the baseline distribution G_0 (details provided in Appendix A.3).

3.3 Simulation study for fixed-p_k BQMR

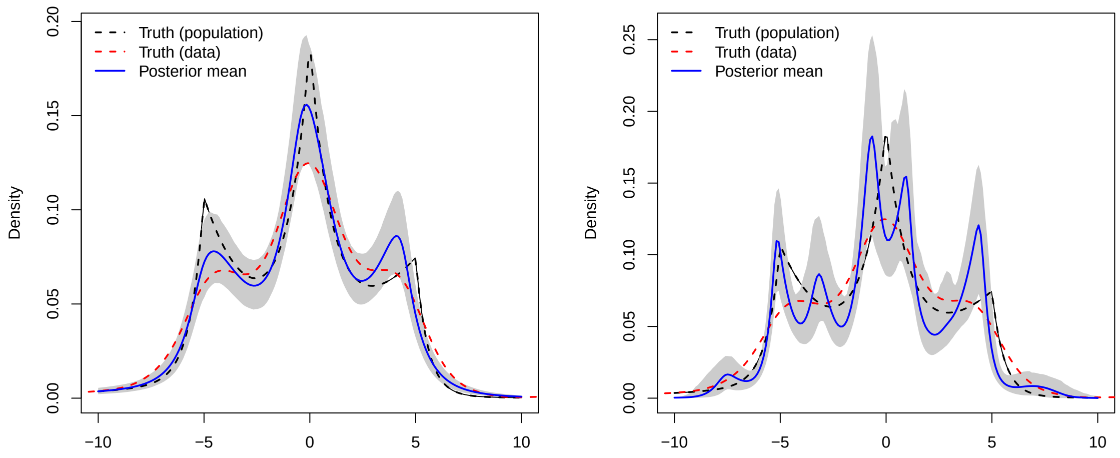
We illustrate the model performance with simulations and consider the same setup as in the simulation study in Section 2.3. Following the standard setting in the literature of linear regression (Tibshirani, 1996; Zou & Yuan, 2008; Li et al., 2010), we simulate data from $y_i = x_i^T \beta + \epsilon_i$, where $\boldsymbol{\beta} = (3, 1.5, 0, 0, 2, 0, 0, 0)^T$. The covariates of each observation are generated independently from a $N_8(0, \Sigma)$, of which the (i, j)th element is $0.5^{|i-j|}$, for $1 \leq i, j \leq 8$. To establish a benchmark for comparison, we define a fixed-p_k AL mixture by formulating the BQMR model with AL_{p_k} components under given $\{p_k\}$. We present results from posterior inference on the predictive error distribution, variable selection and weights, comparing the proposed models with the simpler fixed-p_k AL mixture approach.

In all examples, we estimate the regression vector $\boldsymbol{\beta}$ with a Laplace prior with a Gamma(0.1,0.1) hyperprior for η^2. Independent rescaled-Beta priors are applied for γ_{p_k} with shape parameters $\alpha_\gamma = \beta_\gamma = 3$ and we adopt an inverse-Gamma(2,2) prior for σ. We set $\sigma_\mu^2 = 10$ in the structured priors of $\{\mu_{p_k}\}$. A Dirichlet$(1, \ldots, 1)$ prior is used for $\{\omega_k\}$ in Section 3.3.1 and 3.3.2, while the nonparametric prior is used in Section 3.3.3.

3.3.1 Error density under parametric BQMR

In this section, we demonstrate the advantages of using GAL distributions as the kernel densities with two simulation examples. In the first example, we simulate $n = 600$ observations from the regression setting, where the error ϵ_i follows a weighted mixture of three AL components with $\{p_k\} = \{0.2, 0.5, 0.9\}$, $\{w_k\} = \{0.3, 0.3, 0.4\}$, $\{\mu_{p_k}\} = \{-5, 0, 5\}$ and $\sigma = 0.5$. In the second example, we simulate $n = 100$ data points with a bimodal error distribution, $\epsilon_i \sim 0.5N(-2, 1) + 0.5N(2, 1)$. We fit the data with the proposed BQMR model with $p_k = k/10$ for $k = 1, 2, \ldots, 9$ and the fixed-p_k AL mixture with the same p_k.

Figure 3.1 visualizes the posterior samples of error densities in the first simulation scenario. The posterior mean is marked with the blue line, encompassed by the 95% credible interval (CrI) shaded in gray. The generating density is plotted with a blacked dashed line, while the red line represents the empirical error density in the data. We see that the proposed BQMR model (M_1) captures the truth quite well, while the fixed-p_k AL mixture (M_2) completely misses the true density and produces very jagged predictions.



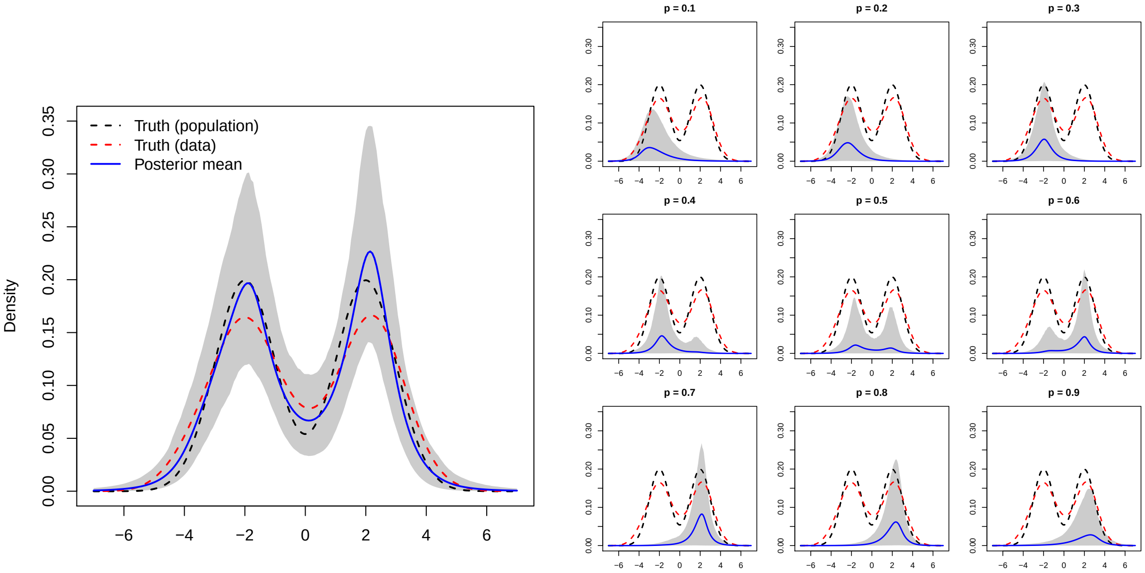
(a) M_1: Fixed-p_k BQMR (GAL kernel)

(b) M_2: Fixed-p_k AL mixture

Figure 3.1: Simulation study comparing fixed-p_k BQMR (GAL kernel) with the fixed-p_k AL mixture: posterior mean and 95% pointwise interval of predictive error density; data generated from mixtures of AL densities

In this example, although the data is generated from a mixture of AL distributions, the BQMR model with GAL components generates predictive error densities that are much closer to the truth than those by the simpler approach with AL components. Instead of identifying the three components that truly participated in the data generation, the fixed-p_k AL approach with $K = 9$ components tries to fit the data with substantial contribution from all components, which is likely the major reason behind its ragged predictive density and poor performance.

Figures 3.2 and 3.3 summarize the results for the second simulation scenario, where the errors arise from two equally weighted normal distributions. Each figure includes the visualization of predictive error density as well as the by-component plot where the posterior mean and 95% CrI bands are visualized for each weighted component. In this simulation setting, the components in the error distribution are Gaussian and do not belong to the



(a) Predictive error density

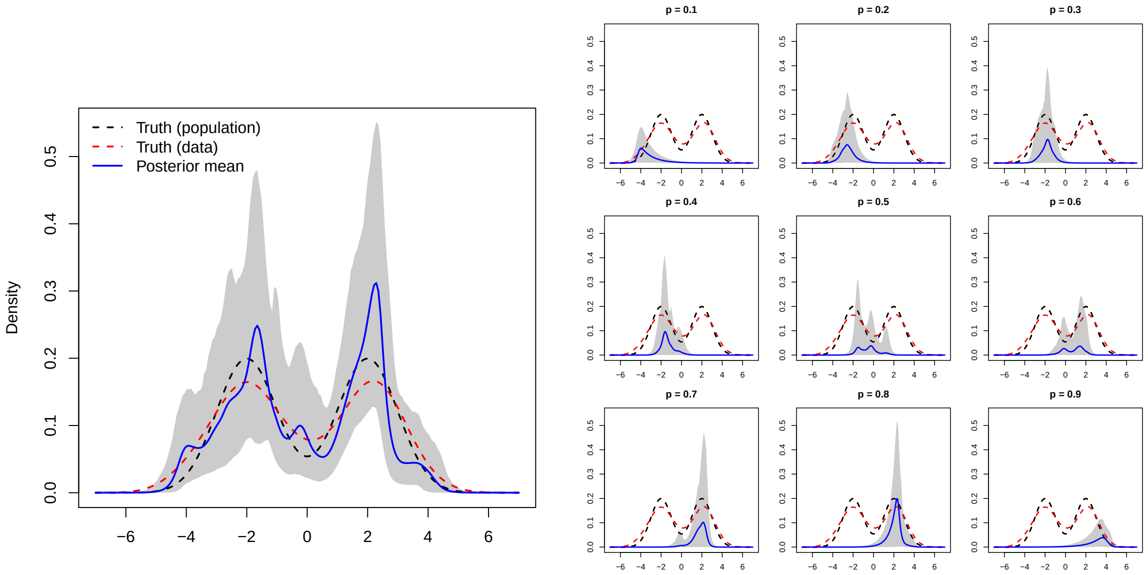
(b) Contribution from weighted components

Figure 3.2: Simulation study of fixed-p_k BQMR with GAL kernel (M_1): posterior mean and 95% pointwise interval of predictive error density (left) and the individual contribution from each weighted component (right); data generated from mixtures of two normals

family of the GAL (or AL) distributions. However, constructed with the flexible GAL components, the BQMR framework (M_1) still fit the data quite well and the predictive error distribution highly resembles the true underlying mixture of normals. The fixed-p_k AL mixture (M_2), on the other hand, struggles in this scenario and produces spiky posterior predictions for the error density, owing to the fact that the AL distributions can only behave in a very restricted way when $\{p_k\}$ are fixed. This simulation study again reflects the importance of kernel density in the BQMR structure.

3.3.2 Identification of influential predictors

One of the major applications of the BQMR framework is in identifying influential predictors, because the procedure summarizes information from multiple parts of the



(a) Predictive error density

(b) Contribution from weighted components

Figure 3.3: Simulation study of fixed-p_k AL mixture (M$_2$): posterior mean and 95% pointwise interval of predictive error density (left) and the individual contribution from each weighted component (right); data generated from mixtures of two normals

response distribution and assesses the predictive effect of the covariates in a comprehensive way. We design a simulation study to illustrate the performance of the proposed BQMR framework in this aspect. The following error distributions are considered:

- ϵ_i follows a Student-t distribution with 1 degree of freedom
- ϵ_i follows a Student-t distribution with 3 degrees of freedom
- $\epsilon_i \sim 0.5N(-2, 1) + 0.5N(2, 1)$

The first error density is chosen to test the model performance when the true distribution has undefined variance. The second illustrates a fat-tailed density, while the last represents a bimodal case. Under each error distribution, we generate $N = 100$ data sets with $n = 100$ observations and fit the following models on each data set:

	Error distribution					
	Student t with df 1		Student t with df 3		Mixture of normals	
Model	TP	FP	TP	FP	TP	FP
BQMR with GAL$_{p_k}$						
M$_1$: $K = 9$	3.00 (0.00)	0.07 (0.26)	3.00 (0.00)	0.24 (0.51)	3.00 (0.00)	0.10 (0.30)
M$_2$: $K = 3$	3.00 (0.00)	0.09 (0.29)	3.00 (0.00)	0.28 (0.55)	3.00 (0.00)	0.13 (0.34)
Fixed-p_k AL mixture						
M$_3$: $K = 9$	2.50 (0.96)	0.12 (0.52)	3.00 (0.00)	0.65 (0.93)	3.00 (0.00)	0.61 (0.93)
M$_4$: $K = 3$	2.53 (1.00)	0.05 (0.22)	3.00 (0.00)	0.97 (1.21)	2.98 (0.14)	2.09 (1.26)

Table 3.1: Mean (standard deviation) of true positives (TP) and false positives (FP)

- M$_1$: BQMR with $K = 9$, $p_k = k/10$ for $k = 1, 2, \dots, 9$
- M$_2$: BQMR with $K = 3$, $p_k = k/10$ for $k = 1, 5, 9$
- M$_3$: Fixed-p_k AL mixture with $K = 9$, $p_k = k/10$ for $k = 1, 2, \dots, 9$
- M$_4$: Fixed-p_k AL mixture with $K = 3$, $p_k = k/10$ for $k = 1, 5, 9$

For each data set, if the 95% highest posterior density interval of β_j does not include zero, then we consider x_j as selected by the model. Based on this criterion, we summarize the number of correctly included predictors (True Positive, abbr. TP) and the number of incorrectly included predictors (False Positive, abbr. NP) across all 100 data sets for each simulation scenario. With the true $\boldsymbol{\beta}$ being $(3, 1.5, 0, 0, 2, 0, 0, 0)$, the optimal values for TP and FP are 3 and 0, respectively.

Table 3.1 summarizes the mean and standard deviation of TP and FP under each error distribution. Compared with the fixed-p_k AL mixture, the proposed BQMR framework has a better (scenario 1 and 3) or equivalent (scenario 2) performance in selecting the truly active predictors. Overall, the number of incorrect inclusions is also lower under the BQMR framework. The first scenario of Student-t_1 errors is the most challenging in that the error distribution does not have a finite variance. In this case, the proposed BQMR models (M_1 and M_2) select the active x_j correctly in all 100 data sets (TP with mean 3 and standard deviation 0), while fixed-p_k AL mixture had quite some difficulty identifying the true contributing predictors with an average of 2.5 true positives. The average false positives over all data sets are comparable across models in this case.

In scenario 2 where the errors follow a Student-t_3 distribution, all the models correctly pick up all three active predictors in all data sets, but the proposed framework (M_1 and M_2) produce much lower false positives compared with the simpler approach with AL components (M_3 and M_4). When the error distribution is bimodal in scenario 3, the BQMR framework, again, shows much better accuracy in FP than the AL mixture, in addition to a perfect inclusion outcome indicated by TP with mean 3. Moreover, the standard deviation of TP and FP are a lot lower under the proposed framework in almost all cases, implying lower uncertainty than the simpler fixed-p_k AL mixture.

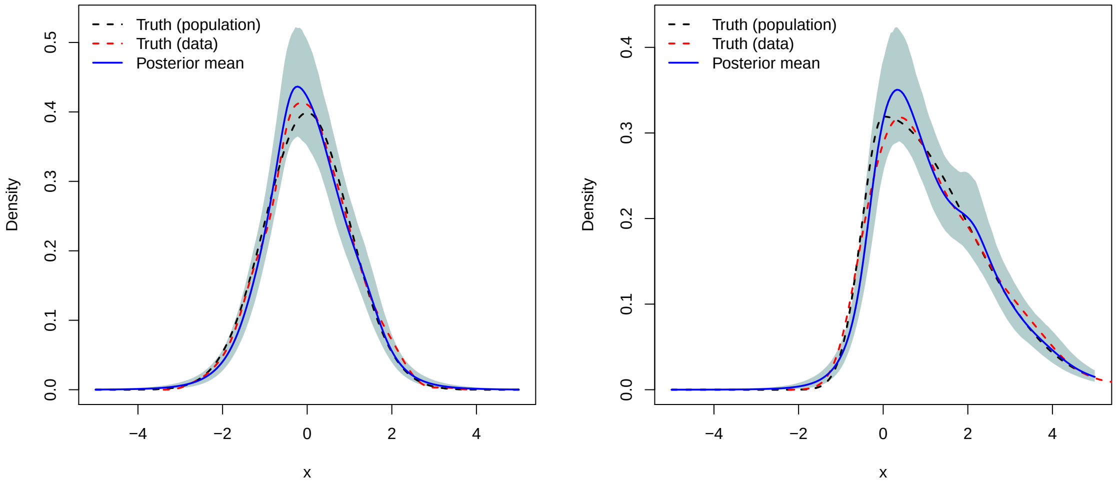
To summarize, the simulation study shows that the proposed BQMR framework with fixed p_k has very high accuracy in terms of identifying both the active predictors and the inactive ones in the linear regression setting. We notice that M_1 and M_2 have quite similar results in all three scenarios, while the only difference between the two models lies in the number of components K. This suggests that in this simulation setting, as few as three GAL components under the proposed framework are adequate for capturing the underlying predictor configuration with a reasonable accuracy.

3.3.3 Extension with nonparametric prior

In this section we simulate data from a skewed error distribution to illustrate the inference of the fixed-p_k BQMR framework with nonparametric prior on weights. We simulate $n = 500$ data points from the regression setting under two scenarios. In the first scenario, we consider a symmetric distribution for the errors and generate ϵ_i independently from a standard normal distribution. The second scenario represents a skewed case in which ϵ_i follows a skew-normal distribution (Fernández & Steel, 1998), where $f(\epsilon) = 2/(\gamma + 1/\gamma)[N(\epsilon/\gamma \mid 0, 1)\mathbb{1}\{\epsilon \geq 0\} + N(\gamma\epsilon \mid 0, 1)\mathbb{1}\{\epsilon < 0\}]$, with $\gamma = 2$. The resulting density has mode at 0 with a concentrated left tail and a fat right tail.

In both cases, we fit the data with the semi-parametric BQMR model with $p_k = k/10$, $k = 1, \dots, 9$. As in Section 3.2.2, a rescaled-Beta$(\mu_G\tau_G, \mu_G(1-\tau_G), 0, p_K)$ applies as the baseline distribution G_0 in the DP prior. The hyperparameters and hyperpriors are chosen to produce a decent amount of variability in the prior of G. A Beta$(12, 12)$ prior is placed on μ_G, so that $\pi(\mu_G)$ is centered at 0.5 with a standard deviation of 0.1. We set the scale $\tau_G = 3$ to favor a unimodal baseline distribution, which corresponds to a G_0 of Beta$(x/p_K \mid 1.5, 1.5)$ when $\mu_G = 0.5$, with most of the probability mass between 0.2 and 0.8. The concentration parameter α_0 determines to what extent the realization from the DP resembles the baseline distribution. We use $\alpha_0 = 10$ so that the marginal prior mean and prior variance of w_k are comparable with those under a Dirichlet$(1, \dots, 1)$ prior.

Figure 3.4 presents the predictive error density under each error distribution based on a single simulated data set. In the left panel, fig. 3.4a shows the mean and 95% CrI band for scenario 1 where the errors follow a standard normal distribution. There is a high



(a) Scenario 1: Normal errors

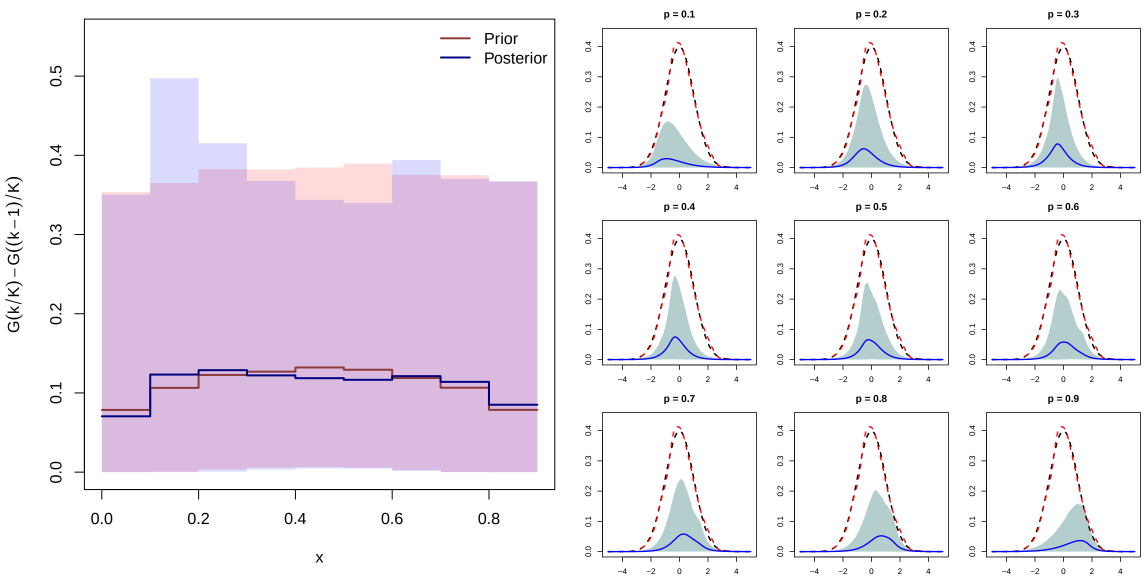
(b) Scenario 2: Skew-normal errors

Figure 3.4: Simulation study of fixed-p_k BQMR with nonparametric prior on the weights: posterior mean and 95% pointwise interval of predictive error density for data sets generated with two different error distributions

resemblance between the model prediction and the generating density of ϵ_i, which suggests a good fit of the framework on the data. We observe similar patterns in the results for scenario 2 (fig. 3.4b). Although in this setting the errors follow a skew normal distribution, the model produces a decent fit and learns the true error density quite well.

Figures 3.5 and 3.6 provide visualizations for the posterior mean (blue line) and 95% interval (light blue bands) of weights $w_k = G(p_k) - G(p_{k-1})$, accompanied by the weighted contribution to the posterior predictive error density from each component. Since the mixture model includes $K = 9$ components, we observe nine steps in the plot of $\{w_k\}$, where the height of the steps from left to right each corresponds to the posterior mean of w_1 to w_K. We overlay the plots with the mean (red line) and 95% intervals (pink band) of w_k generated from the DP prior for comparison.

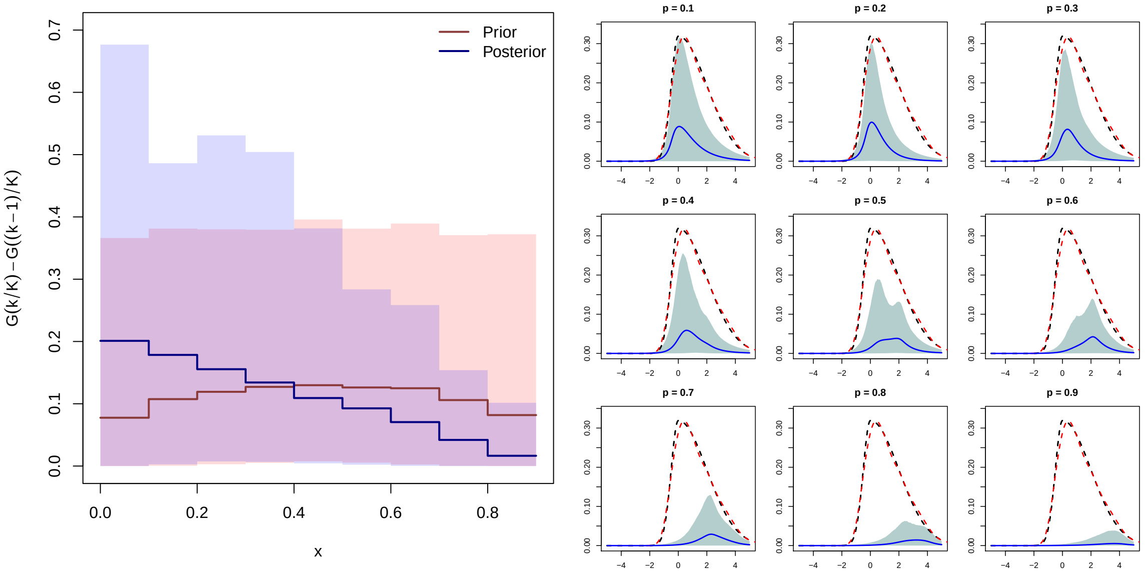
While the prior mean of the weights are similar across components in both fig. 3.5a and fig. 3.6a, the posterior distribution of w_k varies for the two scenarios. In scenario 1 where the error distribution is symmetric, the posterior weights are quite balanced across components, with the lower and upper quantile components weighted slightly less compared with those in the center (Figure 3.5). The fact that the prior and the posterior almost overlap in fig. 3.5a suggests that the model does not learn much in this setting. However, in the second scenario of skewed errors, the model assigns much larger weights in the posterior to the leftmost components (Figure 3.6). This is consistent with the pattern in the true error density that the left tail is more concentrated than the right. Overall, the results suggest that the semi-parametric BQMR framework offers useful posterior inference on the component weights under relaxed assumptions on the prior distribution, particularly when the conditional response distribution is skewed.



(a) Weights for each quantile component

(b) Predictive error density by weighted component

Figure 3.5: Simulation study for nonparametric prior on weights: posterior inference (mean and 95% pointwise interval) on component weights (left) and on the individual contribution to predictive error density from each weighted component (right) under scenario 1; data generated with normal errors



(a) Weights for each quantile component

(b) Predictive error density by weighted component

Figure 3.6: Simulation study for nonparametric prior on weights: Posterior inference (mean and 95% pointwise interval) on component weights (left) and on the individual contribution to predictive error density from each weighted component (right) under scenario 2 ; data generated with skew-normal errors

3.4 Bayesian quantile mixture regression with random p_k

In Section 3.2, we devise the BQMR framework based on the GAL distribution with fixed percentage p_k. Owing to the shape parameter γ in the GAL distribution, all quantile components have varying skewness and shape, which substantially increases the flexibility of the model. Another way of constructing a flexible BQMR is to take the AL densities as mixture kernel, but set the percentage p_k to be random. By allowing p_k to vary, we shift the focus from specifying specific quantiles to estimating the predictive effects given a fixed total number of random components as well as to learning the configuration of the quantile components.

3.4.1 Model formulation and inference

The random probabilities $\{p_k\}$ can be described with a homogeneous Poisson process (HPP). Denote $K \geq 2$ as the number of mixture components. If we consider ordered probabilities $p_1, \ldots, p_K$ as arrival times in the time interval of $(0, 1]$, we can model $\{p_k\}$ using an HPP on the unit interval. By definition, conditional on K, the arrival times $p_1, \ldots, p_K$ follow a uniform distribution with $p_1, \ldots, p_K \propto \mathbb{1}\{0 < p_1 < \ldots < p_k < \ldots < p_K < 1\}$.

The above specification provides a natural prior for $\{p_k\}$ treated as the skewness parameter of the AL components and completes the BQMR framework with random p_k. Given that the AL distribution is a special case of the GAL distribution with shape parameter $\gamma = 0$, we can express the random-p_k BQMR model in the following hierarchical form and place an HPP prior on the probabilities,

$$\begin{aligned}
y_i \mid \boldsymbol{\beta}, \{p_k\}, \{\mu_{p_k}\}, \sigma, v_i, \boldsymbol{\xi}_i &\sim \prod_{k=1}^{K} \Big[N(y_i \mid \mu_{p_k} + \boldsymbol{x}_i^T \boldsymbol{\beta} + A_{p_k} v_i, \sigma B_{p_k} v_i) \Big]^{\xi_{ik}} \\
\boldsymbol{\xi}_i \mid \omega_1, \ldots, \omega_k &\sim Multinomial(\boldsymbol{\xi}_i \mid \omega_1, \ldots \omega_k) \\
v_i \mid \sigma &\overset{\text{i.i.d}}{\sim} Exp(v_i | 1/\sigma) \\
\pi(p_1, \ldots, p_K) &\propto \mathbb{1}\{0 < p_1 < \ldots < p_k < \ldots < p_K < 1\}
\end{aligned}$$

The posterior sampling scheme of this BQMR model is very similar to that under the parametric BQMR with fixed p_k in Section 3.2. If we define $p_0 = 0$ and $p_{K+1} = 1$, we can write the full posterior of p_k given all other parameters as,

$$p(p_k \mid \ldots) \propto \prod_{i=1}^{n} \Big[N(y_i \mid \mu_{p_k} + \boldsymbol{x}_i^T \boldsymbol{\beta} + A_{p_k} v_i, \sigma B_{p_k} v_i) \Big]^{\xi_{ik}} \mathbb{1}\{p_{k-1} < p_k < p_{k+1}\}$$

Note that here A_{p_k} and B_{p_k} are both nonlinear functions of p_k, thus a series of Metropolis-Hastings steps are needed. At iteration m, by applying a truncated normal distribution over

$\left(p_{k-1}^{(m)}, p_{k+1}^{(m-1)}\right)$ as the proposal distribution, we can sample p_k for $1 \leq k \leq K$ sequentially with independent adaptive Metropolis-Hastings steps (details provided in Appendix A.4).

3.4.2 Extension to random K

For the more general version of the model, we make the total number of components K also a random variable to incorporate uncertainty in K. Considering that it is not practical to implement a BQMR model with an infinite number of components, we cap K from above with a reasonable upper bound K_{max}. Also, given the model structure and objectives, it makes sense to think of a K_{max} even for problems where one knows very little about the distribution.

One way of selecting K is to we can put a prior on $\{K : K \leq K_{max}, K \in \mathbb{N}\}$ and fit a BQMR model for each K. Trans-dimensional MCMC algorithms, such as reversible jump MCMC by Green (1995), can potentially be used to evaluate the posterior distribution of K. However, practically it would be difficult to implement the method with satisfying acceptance rate of the Metropolis-Hastings steps, given the fact that the model already involves a series of Metropolis-Hastings steps for sampling $\{p_k\}$. Therefore, we consider approximating the posterior model probability following the method in Scott (2002).

Moreover, we believe that a reasonable finite upper bound for K is desirable in this setting, because in BQMR framework, each component has a specific meaning associated with the corresponding p_k. Realistically, we would not expect K to be very large in most cases. Given the objectives and the concept of the framework, posterior model probability on a constrained parameter space of K can provide useful information for inferences on the number of components.

Consider $\boldsymbol{\phi}_K = (\boldsymbol{\beta},\, \boldsymbol{p}_{1:K},\, \boldsymbol{\mu}_{p_{1:K}},\, \sigma)$ and $\boldsymbol{\phi} = \{\boldsymbol{\phi}_1, \ldots, \boldsymbol{\phi}_{K_{max}}\}$, the full parameter vector for a particular K. As in Scott (2002), we calculate the approximate posterior model probability with M posterior samples drawn with separate MCMC algorithms given K:

$$\begin{aligned}
p(K \mid \boldsymbol{y}, \boldsymbol{x}) &= \int p(K \mid \boldsymbol{y}, \boldsymbol{x}, \boldsymbol{\phi}) p(\boldsymbol{\phi} \mid \boldsymbol{y}, \boldsymbol{x}) \mathrm{d}\boldsymbol{\phi} \\
&\approx \frac{1}{M} \sum_{j=1}^{M} p(K \mid \boldsymbol{y}, \boldsymbol{x}, \boldsymbol{\phi}^{(j)}) \\
p(K \mid \boldsymbol{y}, \boldsymbol{x}, \boldsymbol{\phi}^{(j)}) &\propto p(\boldsymbol{y} \mid K, \boldsymbol{x}, \boldsymbol{\phi}_K^{(j)}) p(K) \\
&\propto \left\{ \prod_{i=1}^{n} \left[\sum_{k=1}^{K} w_k^{(j)} AL_{p_k^{(j)}} \left(y_i \,\middle|\, \mu_{p_k}^{(j)} + \boldsymbol{x}_i^T \boldsymbol{\beta}^{(j)}, \sigma^{(j)} \right) \right] \right\} p(K)
\end{aligned}$$

With little prior knowledge about K, we complete the framework with a flat prior on the number of components, such that $p(K) \propto \mathbb{1}\{k \leq K_{max}\}$, $K \in \mathbb{N}$.

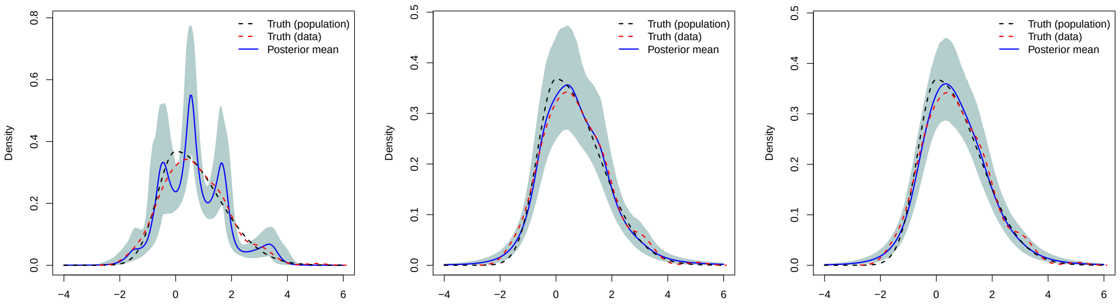
For large data sets, the posterior model probability may favor models with a large number of mixture components (large K), not only because the data carry enough information to support the estimation of a good number of components, but also because models with larger K tend to produce a better fit to the local behavior of the response distribution. In this circumstance, we may prefer priors that penalize very large K to reach a balance between parsimonious structure and goodness-of-fit.

3.5 Simulation study for random-p_k BQMR

In this simulation study, we shift the focus from identification of influential predictors to the estimation of $\{p_k\}$. We consider the linear regression setting $y = \boldsymbol{x}^T\boldsymbol{\beta} + \epsilon$ with two predictors $x_1, x_2 \sim_{\text{i.i.d.}} N(0, 1)$, with $\boldsymbol{\beta} = (4, -3)^T$ and the error distribution specified later in this section. Since $\boldsymbol{\beta}$ is not sparse, the posterior inference for the regression coefficients will be similar between a Gaussian prior and a Laplace prior on $\boldsymbol{\beta}$. In all cases, we fit a K-component random-p_k BQMR model with known K and use the same priors as in Section 3.3 whenever applicable. For the weight parameters $\{\omega_k\}$, we use a Dirichlet(1,...,1) prior.

To assess the performance of the random-p_k framework, we first simulate the data from a weighted mixture of two AL distributions with known parameter values. The data set can be viewed as a realization from the random-p_k framework with point mass probability on percentage p_k of each component. For each data set, we generate $n = 200$ observations and fit the data with the random-p_k BQMR discussed in 3.4.1 (results not shown). When the components are reasonably far apart, the proposed model can identify the components and estimate the accompanying p_k quite well if the total number of components K is correctly specified. When the data is fitted with more mixture components than there actually is, in addition to the truly active components, the model may pick up a few extra components. Nevertheless, all components combined produces a posterior predictive error density that highly resembles the underlying true error density, which is captured in the 95% credible band of posterior samples.

We consider again the skew-normal distribution (Fernández & Steel, 1998) for ϵ_i, as in Section 3.3.3, to illustrate the scenario where the errors do not arise from a mixture distribution. With $\gamma = 1.5$ and $\sigma = 1$, the generating distribution of ϵ_i is right-skewed with mode at 0. We simulate $n = 200$ observations and fit the data with the following models for comparison: a fixed-p_k AL mixture with $\{p_k\} = \{0.10, 0.25, 0.50, 0.75, 0.90\}$ (M$_1$), a five-component BQMR model with random p_k (M$_2$), followed by a fixed-p_k BQMR model



(a) M_1: AL, fixed p_k (b) M_2: AL, random-p_k $(K = 5)$ (c) M_3: GAL, fixed-p_k

Figure 3.7: Simulation study comparing fixed-p_k BQMR with different kernel specification and random-p_k BQMR: posterior mean and 95% pointwise interval of error density; data generated with skew-normal errors

with GAL kernels and same $\{p_k\}$ as in the fixed-p_k AL approach (M_3).

Comparing the results from fig. 3.7a (M_1) and fig. 3.7b (M_2) , we see that the random-p_k framework fits the data much better and produces a smoother posterior predictive error density than the fixed-p_k AL mixture. The 95% CrI of M_2 captures most of the behavior of ϵ_i, while M_1, the simpler approach, restricted by the unadjustable skewness of AL kernels under fixed p_k, struggles to emulate the underlying true error density. Results from fig. 3.7b (M_2) and fig. 3.7c (M_3) are similar, with M_3 providing an even smoother fit and narrower posterior intervals for the density estimate. This suggests that for this data set, with GAL kernels the fixed-p_k BQMR framework exhibits more flexibility than the random-p_k model with AL components and shows the best fitting among all three candidate models. Figure 3.8 presents the by-component posterior inference of M_2 and shows the posterior mean of p_k and w_k of each component. The first three components all have $p_k < 0.5$ and they receive much heavier weights than the two components in the right tail, which indicates that the left-to-center of the distribution plays a more important role in the regression analysis.

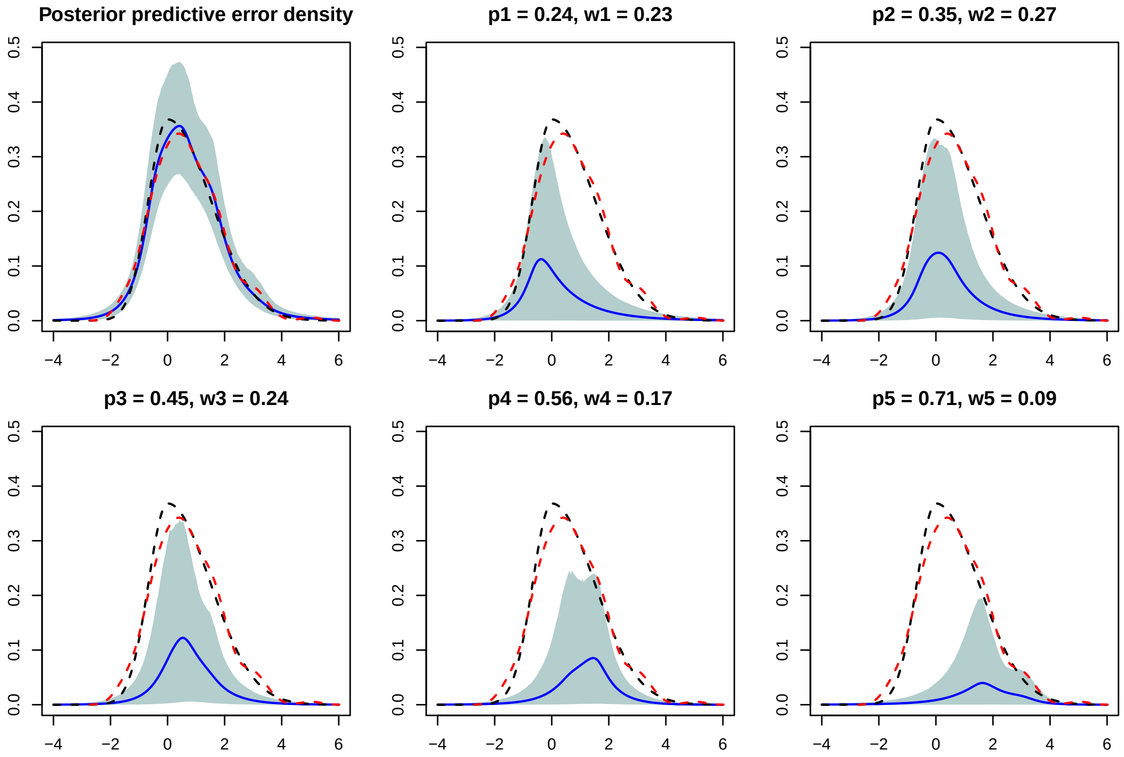


Figure 3.8: Simulated skewed-normal data fitted with random-p_k BQMR (M_2): Posterior mean and 95% pointwise interval of predictive error density (top left) and of weighted contributions from component k labeled with posterior mean of p_k and w_k

Finally, we fit the BQMR framework on the same $n = 200$ observations with skew-normal errors to illustrate the posterior inference for random number of mixture components K in Section 3.4.2. Given that the sample size is 200, we consider $K_{max} = 19$ mixture components as the upper bound and put a uniform prior on K, such that $\pi(K) \propto 1$ for $K \in \{2, 3, \ldots, 19\}$. In total, we fit eighteen random-p_k BQMR models and calculate the approximate posterior model probability following the method in Section 3.4.2. As the number of the components K increases, the posterior model probability keeps increasing and culminates at $K = 10$. Afterwards, the probability slowly drops for all K greater than ten (capped at nineteen, Figure 3.9). The inference on the approximate posterior model probability suggests that under the BQMR model with AL kernel and random p_k, a total number of components $K = 10$ is preferred for modeling this data set. However, some among these ten components can be potentially lumped together based on the proximity of the posterior estimates of the percentiles p_k to achieve a more parsimonious model.

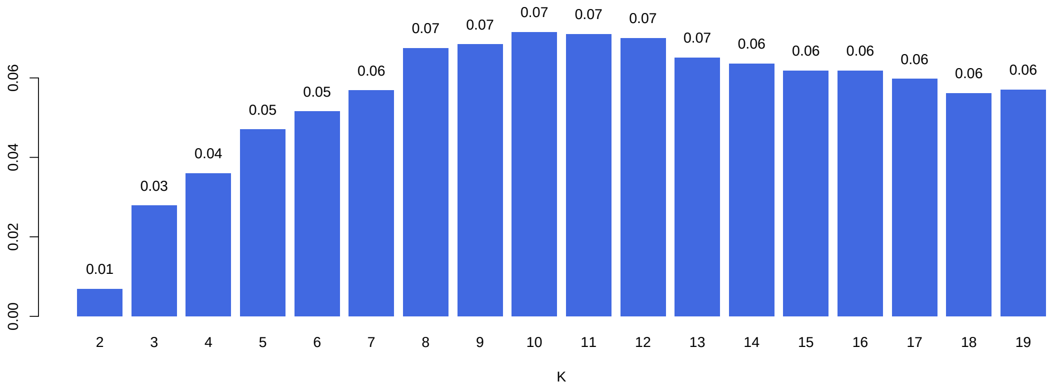


Figure 3.9: Simulated skewed-normal data fitted with random-p_k BQMR with $K_{\max} = 19$: Approximate posterior model probability by K

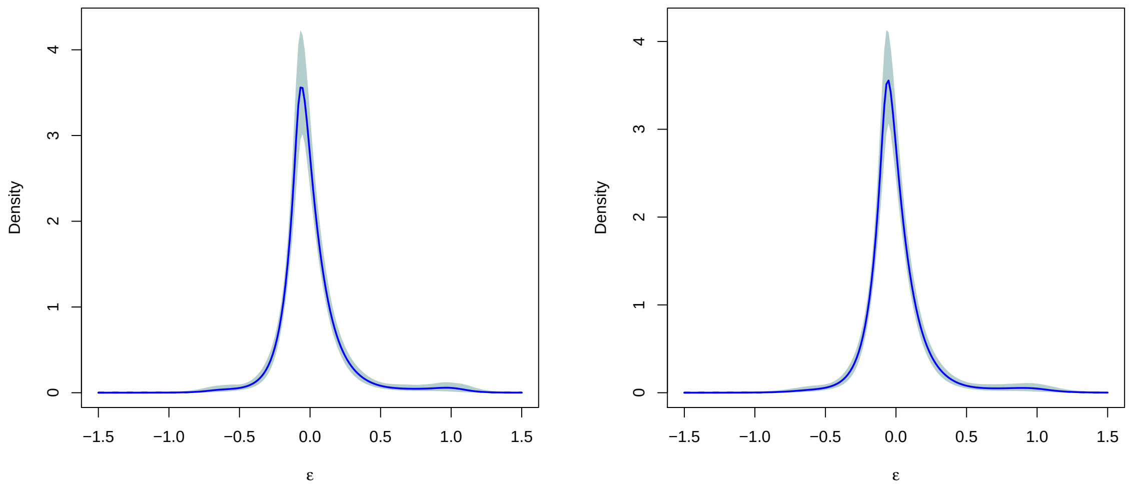
3.6 Data example: Boston housing data

We illustrate and compare the proposed framework with again the realty price data from the Boston Standard Metropolitan Statistical Area (SMSA) in 1970 (Harrison & Rubinfeld, 1978), which contains $n = 506$ observations. The response variable is the log-transformed corrected median value of owner-occupied housing in USD 1000 (LCMEDV), and we include the same fifteen predictors as in the analysis in Section 2.4.2.

Using the proposed BQMR framework, we approach the identification of important variables through modeling the conditional response distribution. We consider the following BQMR models for comparison: random p_k with AL kernels (M_1); and fixed p_k with GAL components (M_2). We use the same priors for all parameters common to both models, including a Laplace prior for $\boldsymbol{\beta}$ with Gamma(0.1,0.1) as the hyperprior for η^2, an inverse-Gamma(2,2) prior for σ, truncated normal priors with $\sigma_\mu^2 = 10$ for intercepts μ_{p_k} and a Dirichlet(1,...,1) for the weights $\{\omega_k\}$. As introduced in Section 3.4.1, for the random-p_k approach, we assume that $\{p_k\}$ follows a homogeneous Poisson process a priori. As for the fixed-p_k model, we apply independent rescaled-Beta priors with shape parameters $\alpha_\gamma = \beta_\gamma = 3$ for γ_k to indicate a slight preference for values that are not too close to the boundaries L_k or U_k.

We begin the analysis by considering $K = 9$ components for both models. In the fixed-p_k approach, we set $p_k = k/10$ for $k = 1, \ldots, 9$. The posterior inference for predictive error density and the by-component contribution under each model are presented in Figures 3.10 and 3.11. Both models predict a slightly right-skewed error density with heavy tails and most of the probability mass between -0.5 and 0.5. The posterior predictive error densities produced by the two models turn out to be very similar. There is also a lot of resemblance between the by-component plots of the two models (Figure 3.11). In both analyses, the third, fourth and the fifth component contribute the most to the predictive error density, with the posterior mean of the corresponding p_k being 0.31, 0.39 and 0.46 under M_1 and the fixed p_k equal to 0.3, 0.4 and 0.5 under M_2. This suggests that the left-to-center of the conditional response distribution plays a more important role and is likely to be more sensitive to changes in the predictors.

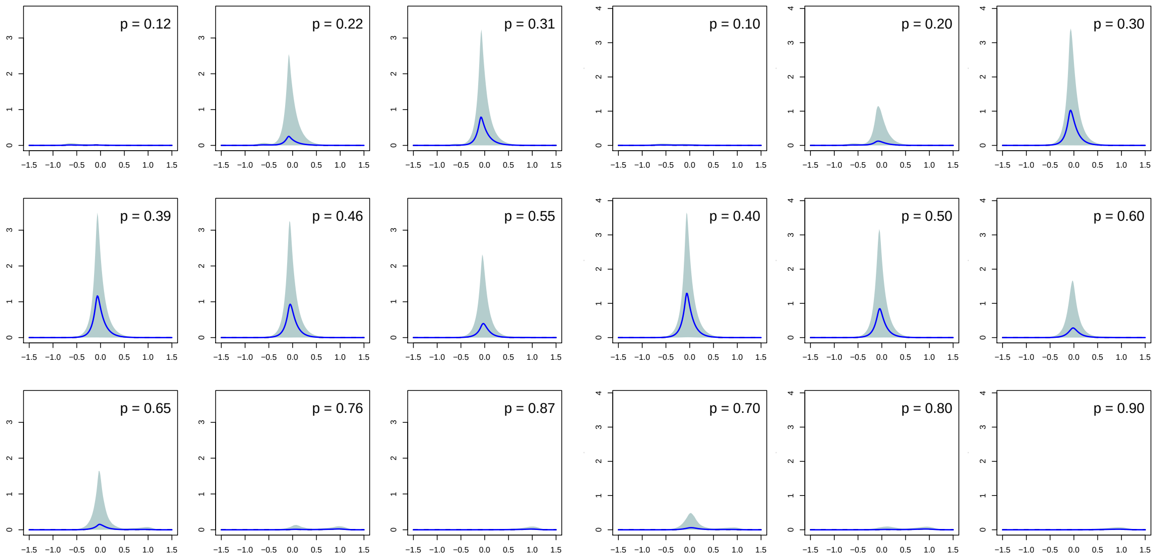
In the by-component analysis in Figure 3.11, the first and the last component receive very low weights in both models, which implies that a more parsimonious model with fewer than nine components can be fitted for this data. We carry out the posterior



(a) M_1: random p_k, AL

(b) M_2: fixed p_k, GAL

Figure 3.10: Boston housing data: Comparison of random-p_k BQMR with fixed-p_k BQMR: Posterior mean and 95% pointwise interval of predictive error density



(a) M_1: random p_k, AL (label: posterior mean p_k)

(b) M_2: fixed p_k, GAL

Figure 3.11: Boston housing data: Comparison of random-p_k BQMR with fixed-p_k BQMR: Posterior mean and 95% pointwise interval of individual contributions to the error density by weighted components

model probability inference with respect to the space of K capped at $K_{max} = 9$. Under a uniform prior, $K = 7$ is associated with a higher posterior probability than the other values.

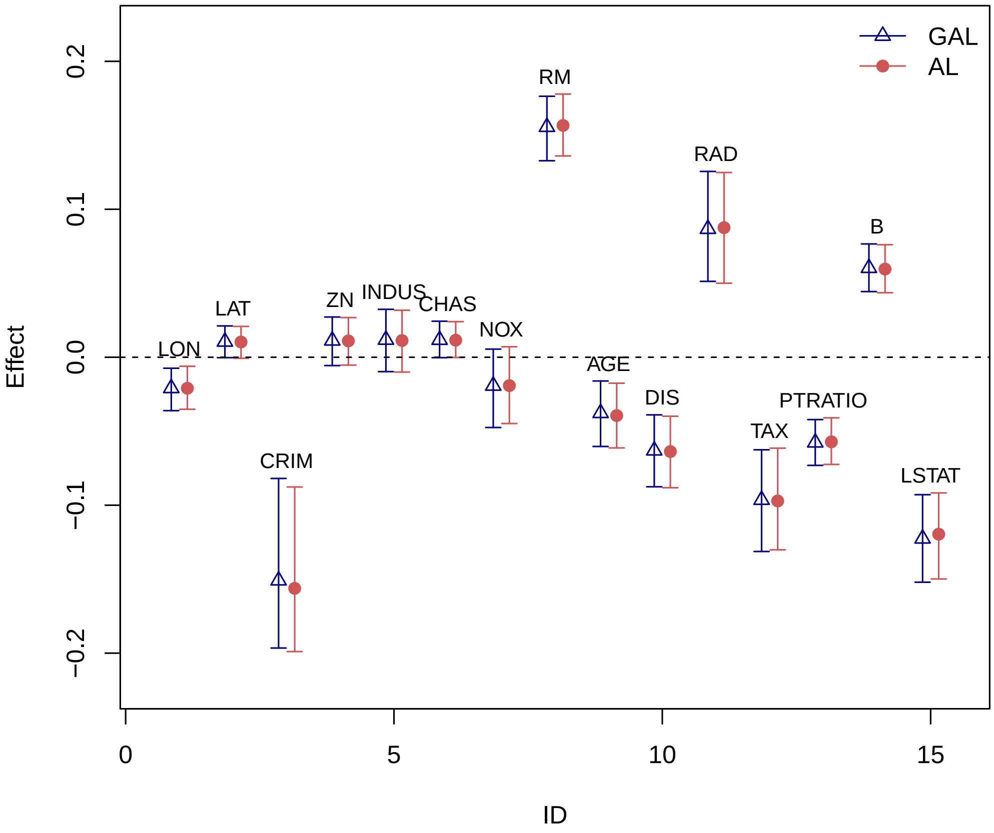


Figure 3.12: Boston housing data: Comparison of random-p_k BQMR with fixed-p_k BQMR: Posterior mean and 95% HPD interval of β_j , $j = 1, \ldots, 16$ under M$_1$ (random-p_k, AL kernel) and M$_2$ (fixed-p_k, GAL kernel)

Figure 3.12 shows that interestingly, the two models produce very similar posterior inference on the regression coefficients. The posterior mean and 95% highest posterior density (HPD) interval of all variables are almost the same for the two approaches. Since the early-on inference on the number of components to be included suggests that $K = 7$ is adequate for a decent fit on the data, we revise both models to include only seven components and specify $p_k = (k + 1)/10$, $k = 1, \ldots, 7$ for the fixed-p_k approach. The results are very similar to those of models with $K = 9$ (Figure 3.12). Except LAT, ZN, INDUS, CHAS and NOX, the 95% HPD intervals of the effects of all remaining predictors do not include zero under either M_1 or M_2. Per capita crime (CRIM) has the largest negative impact on the realty price, while average number of rooms per dwelling (RM) boosts up the property value the most. The original purpose of this data set was to analyze whether the housing price was associated with the air quality evaluated with the nitric oxides concentration (parts per 10 million) per town (NOX) (Harrison & Rubinfeld, 1978). Although NOX has a negative effect on the response, the 95% posterior interval of its regression coefficient includes zero, indicating that the realty prices are overall insensitive to the air quality.

To assess the model performance, we calculate the log-pseudo-marginal-likelihood (LPML) (Geisser & Eddy, 1979; Gelfand et al., 1992; Gelfand & Dey, 1994). Geared to prediction, LPML is a leave-one-out (n-fold) cross-validation measure with log likelihood as the criterion, $LPML = \sum_{i=1}^{n} \log(CPO_i)$, where $CPO_i = f(y_i \mid \boldsymbol{y}_{-i})$ is the conditional predictive ordinate and $\boldsymbol{y}_{-i}$ is the response vector without y_i. Using the M posterior samples from the Markov chain, an estimate of CPO can be obtained with $\widehat{CPO_i} = \left\{ \frac{1}{M} \sum_{m=1}^{M} \left[f(y_i \mid \boldsymbol{\psi}^{(s)}) \right]^{-1} \right\}^{-1}$, where $\boldsymbol{\psi}^{(m)}$ stands for the mth posterior sample of all model parameters. Larger LPML values represent better model-fitting from a predictive standpoint. For comparison, in addition to BQMR, we fit five Bayesian quantile regression (BQR) models with comparable lasso prior on the regression coefficients for the 10th, 30th, 50th, 70th and 90th quantile under both the AL and the GAL distribution.

Table 3.2 summarizes the LPML by different models. With $K = 7$ components, the fixed-p_k BQMR with GAL kernel and the random-p_k BQMR with AL kernel achieve the highest LPML values among all models considered, with the latter being slightly higher. This result indicates a better fit for the Boston housing data with the BQMR framework than with BQR. Among the single quantile regressions, the 30th and the 50th percentile models have higher LPML than those for the other quantiles. This coincides with our finding from the BQMR analysis that the left-to-center of the conditional response distribution appears to play a more important role in regression and prediction.

Kernel/error		*BQR*				
specification	*BQMR,* $K = 7$	$p = 0.1$	$p = 0.3$	$p = 0.5$	$p = 0.7$	$p = 0.9$
GAL	193.8	163.6	179.3	178.3	166.9	117.4
AL	195.1	99.9	182.4	175.5	117.4	-37.1

Table 3.2: Boston housing data: Log-pseudo-marginal-likelihood of fixed-p_k and random-p_k BQMR with seven components and BQR with GAL and AL errors for the 10th, 30th, 50th, 70th and 90th quantile

The proposed BQMR framework offers a practical way to analyze data sets where the conditional distribution of the responses is non-normal. In the example of Boston housing data, both the random-p_k BQMR method and the fixed-p_k approach produce a posterior predictive error density that is slightly skewed to the right, with both tails heavier than those of a Gaussian distribution. Analyzing such data set with some standard regression procedures may lead to inadequate fitting and unsatisfactory estimation for the predictor effects. The mixture framework we propose provides a flexible alternative to capture the non-normal pattern in the errors. Further, since the BQMR model consists of components parameterized by quantiles, it presents intuitive results on which parts of the conditional response distribution contribute more to the formulation of the response. In the Boston housing example, more contributions are observed for the mixture components of between the 20th and 70th percentile in the BQMR analysis. This type of findings can easily be obtained in the proposed quantile mixture regression framework.

3.7 Discussion

We have developed a Bayesian mixture quantile regression framework to integrate information from multiple parts of the response distribution to inform the estimation of the regression coefficient. The mixture components are parameterized in terms of quantiles and all components share the same regression coefficient vector. We devise the framework under two choices of kernel densities: the GAL distribution and the AL distribution. The former applies to the scenario when we have a list of percentiles we are interested in, while the later can be used to estimate the percentiles that play a more important role in the regression. The two versions of the framework tend to produce similar inference on the regression coefficients, with the fixed-percentile model showing slightly lower predictive uncertainty owing to the smoothness and the flexibility of the GAL densities. Compared with single quantile regressions, the BQMR models show better model-fitting from a predictive perspective in the Boston housing data example. The hierarchical structure of the kernel densities makes it straightforward to implement the model and draw posterior inference with Markov chain Monte Carlo methods.

The main motivation for this work as well as our main contribution is to develop a regression procedure that considers comprehensively the effect of predictors on different parts of the response distribution. The idea resonates with the key purpose of the classical composite quantile regression. Instead of focusing on a summary statistic of the response distribution, such as expectation in mean regression and a certain quantile in simple quantile regression, the proposed mixture framework forms posterior inferences based on the big picture of the conditional response distribution. Simulation studies show that the model has good performance and great potential in both identification of influential variables and prediction, especially when the errors follow a multi-modal or heavy-tailed distribution. The mixture model provides a convenient tool to analyze observations arising from a complicated data generating mechanism as well as to answer study questions involving several regions of the response distribution.

Chapter 4

Modeling and inference for survival analysis and ROC curve estimation

In this chapter, we explore applications of BQMR to two popular and important research topics in biomedical sciences and epidemiology: survival analysis and estimation of receiver operating characteristics (ROC) curve. In the setting of a controlled study, both survival analysis and ROC curve estimation involve the following two arms: the controlled group and the actively-treated group (or diseased group in ROC estimation). The challenge lies in the fact that the response distribution of the two cohorts can vary quite a lot from each other. In this case, using a common error distribution for both cohorts is restrictive.

Our contribution is established on fitting the two cohorts with flexible BQMR that allows for each cohort to have its own response distribution; yet carefully constructed so that the modeling framework conforms with the underlying assumption and generating mechanism of the data, for instance, right-censoring in survival data sets. Fitting the two arms with separate BQMR framework offers insights on the relative importance of different sections in the distribution of the response variable of each arm, that is, survival time and test score for disease diagnosis, as well as on the cohort-specific identification of important variables. For ROC estimation without predictors, we further develop a two-cohort BQMR framework to model both arms simultaneously. The framework is designed with care to allow information sharing across the two arms and to ensure stochastic ordering in the response distribution, a biologically plausible assumption in certain applications.

4.1 Quantile mixture regression for survival analysis problems

4.1.1 Background

Survival analysis is concerned with inference and prediction for time-to-event data, a.k.a. survival time data, with the event commonly defined as death of patient or failure of disease management. Oftentimes patients are randomized into two cohorts to receive different treatments. Survival time is recorded for each patient and used to compare the effectiveness of the treatments under appropriate model.

Right-censoring commonly exists in survival data, where the censored observations pose technical challenges to estimating the effect of covariates on the likelihood of survival. Consider n survival time and censoring indicator pairs (t_i, δ_i) with associated covariates $\boldsymbol{x}_i$, where $\delta_i = 1$ indicates right-censoring and zero otherwise, then we observe $t_i = \min(t_i^*, c_i)$ and $\delta_i = \mathbb{1}\{t_i^* > c_i\}$, where t_i^* is the latent failure time and c_i is the censoring time.

A popular parametric regression model for survival analysis is the accelerated failure-time (AFT) model. It assumes that a covariate accelerates or decelerates the time to failure by a constant (effect size) (Prentice et al., 1978). Under the AFT model, the hazard function $\lambda(t \mid \boldsymbol{x})$ can be modeled with,

$$\lambda(t \mid \boldsymbol{x}) \quad = \quad \lambda_0(t \exp\{-\boldsymbol{x}^T \boldsymbol{\beta}\}) \exp\{-\boldsymbol{x}^T \boldsymbol{\beta}\}$$

where $\boldsymbol{x}$ is the time-independent covariate vector, $\boldsymbol{\beta}$ is the vector of regression coefficients and $\lambda_0(\cdot) > 0$ is the baseline hazard. Then the following holds for the survival function: $S(t \mid x) = S_0(t \exp\{-\boldsymbol{x}^T \boldsymbol{\beta}\})$, with $S_0(\cdot)$ being the baseline survival when all covariates are zero. This equation indicates that the moderated survival time t given covariates $\boldsymbol{x}$ follows the same distribution as the baseline survival of $t \exp\{-\boldsymbol{x}^T \boldsymbol{\beta}\}$. Consequently, the log-survival time can be expressed as

$$\log(t) \quad = \quad \boldsymbol{x}^T \boldsymbol{\beta} + \epsilon$$

where ϵ_i is distributed as the baseline log-survival time and can be modeled with some distribution, common choices of which are log-normal and log-logistic. We model the logarithmic failure time on the augmented data space with $y_i = \log(t_i^*)$, where $y_i = \log(t_i)$ if $\delta_i = 0$ (fully-observed) and $y_i \geq \log(t_i) = \log(c_i)$ if $\delta_i = 1$ (censored).

4.1.2 Model formulation and implementation

We modify the BQMR framework to fit the AFT models on the augmented data space. Denote $\boldsymbol{y} = (\boldsymbol{y}^o, \boldsymbol{y}^c)$, where $\boldsymbol{y}^o = (\log t_1, \ldots, \log t_{n-m})$ are the log survival times for $n - m$ uncensored data points and $\boldsymbol{y}^c = (\log t_{n-m+1}, \ldots, \log t_n)$ are the log time-at-censoring for m censored observations. For ease of notation, let $\boldsymbol{\theta} = (\boldsymbol{\omega}, \boldsymbol{\mu}, \gamma, \sigma)$. Under a

K-component fixed-p_k BQMR framework, the likelihood given $\boldsymbol{y}$ can be written as,

$$\begin{aligned}\mathcal{L}(\boldsymbol{\beta}, \boldsymbol{\theta} \mid \boldsymbol{y}^o, \boldsymbol{y}^c) &= \prod_{i=1}^{n-m}\left[\sum_{k=1}^{K} w_k GAL_{p_k}(y_i^o \mid \mu_{p_k} + \boldsymbol{x}_i^T\boldsymbol{\beta}, \gamma_{p_k}, \sigma)\right] \\ &\quad \prod_{j=n-m+1}^{n}\left[\int_{y_j^c}^{+\infty}\sum_{k=1}^{K} w_k GAL_{p_k}(y_j^* \mid \mu_{p_k} + \boldsymbol{x}_j^T\boldsymbol{\beta}, \gamma_{p_k}, \sigma)\mathrm{d}y_j^*\right]\end{aligned}$$

By exchanging the integral and the summation, we can express the likelihood given $\boldsymbol{y}^c$ as,

$$\mathcal{L}(\boldsymbol{\beta}, \boldsymbol{\theta} \mid \boldsymbol{y}^c) = \prod_{j=n-m+1}^{n}\sum_{k=1}^{K}\left[\int_{y_j^c}^{+\infty} w_k GAL_{p_k}(y_j^* \mid \mu_{p_k} + \boldsymbol{x}_j^T\boldsymbol{\beta}, \gamma_{p_k}, \sigma)\mathrm{d}y_j^*\right]$$

Then the augmented likelihood given $\boldsymbol{y}$ takes the following form,

$$\begin{aligned}\mathcal{L}(\boldsymbol{\beta}, \boldsymbol{\theta} \mid \boldsymbol{y}, \boldsymbol{y}^*, \boldsymbol{\xi}) &= \prod_{i=1}^{n-m}\prod_{k=1}^{K}\Big[w_k GAL_{p_k}(y_i^o \mid \mu_{p_k} + \boldsymbol{x}_i^T\boldsymbol{\beta}, \gamma_{p_k}, \sigma)\Big]^{\xi_{ik}} \\ &\quad \prod_{j=n-m+1}^{n}\prod_{k=1}^{K}\Big[w_k GAL_{p_k}(y_j^* \mid \mu_{p_k} + \boldsymbol{x}_j^T\boldsymbol{\beta}, \gamma_{p_k}, \sigma)\mathbb{1}\{y_j^* \geq y_j^c\}\Big]^{\xi_{jk}}\end{aligned}$$

where $\mathbb{1}\{\cdot\}$ is the binary indicator function, ξ_{ik} and ξ_{jk} are the auxiliary indicators for the fully observed responses and the censored observations, respectively, which together make up $\boldsymbol{\xi}$, an n-by-k allocation matrix. Finally, by resorting to the hierarchical representation of the GAL kernels, we can express the full posterior in a tractable form,

$$\begin{aligned}\boldsymbol{\beta}, \boldsymbol{\theta} \mid \boldsymbol{y}, \boldsymbol{y}^*, \boldsymbol{\xi}, \boldsymbol{v}, \boldsymbol{s} &\propto \prod_{i=1}^{n-m}\prod_{k=1}^{K}\Big[w_k N(y_i^o \mid \mu_{p_k} + \boldsymbol{x}_i^T\boldsymbol{\beta} + \sigma C_{p_k}|\gamma_{p_k}|s_i + A_{p_k}v_i, \sigma B_{p_k}v_i)\Big]^{\xi_{ik}} \\ &\quad \prod_{j=n-m+1}^{n}\prod_{k=1}^{K}\Big[w_k N(y_j^* \mid \mu_{p_k} + \boldsymbol{x}_j^T\boldsymbol{\beta} + \sigma C_{p_k}|\gamma_{p_k}|s_j + A_{p_k}v_j, \sigma B_{p_k}v_j) \\ &\quad \mathbb{1}\{y_j^* \geq y_j^c\}\Big]^{\xi_{jk}}\prod_{i=1}^{n}\mathrm{Exp}(v_i \mid \sigma^{-1})N^+(s_i \mid 0, 1)\pi(\boldsymbol{\theta})\end{aligned}$$

where A_{p_k}, B_{p_k} and C_{p_k} are the corresponding A, B, C functions of the k-th component and $\pi(\boldsymbol{\theta})$ is the prior of the parameters.

The benefit of constructing the model over the augmented data space is most evident in the convenience of posterior sampling. In fact, the sampling algorithm of the above framework highly resembles that of a standard BQMR, except for an extra update of y_i^* from a truncated normal distribution over $[y_i^c, +\infty)$ centered at $\mu_{p_k} + \boldsymbol{x}_j^T\boldsymbol{\beta} + \sigma C_{p_k}|\gamma_{p_k}|s_j + A_{p_k}v_j$ with scale $\sigma B_{p_k}v_j$.

Inferences for the survival function can be obtained easily based on the posterior samples of parameters. Given $\boldsymbol{\theta}$, the survival function is defined as,

$$\begin{aligned} S(y \mid \boldsymbol{x}, \boldsymbol{\theta}) &= 1 - \int_{-\infty}^{y} \left[\sum_{k=1}^{K} w_k GAL_{p_k}(t \mid \mu_{p_k} + \boldsymbol{x}^T\boldsymbol{\beta}, \gamma_{p_k}, \sigma) \right] \mathrm{d}t \\ &= 1 - \sum_{k=1}^{K} w_k F^{GAL_{p_k}}(y \mid \mu_{p_k} + \boldsymbol{x}^T\boldsymbol{\beta}, \gamma_{p_k}, \sigma) \end{aligned} \tag{4.1}$$

where $F^{GAL_{p_k}}(y \mid \mu, \gamma, \sigma)$ stands for the cumulative distribution function of a GAL_{p_k} distribution at y with location μ, shape γ and scale σ, which is in closed form (Section 2.1.2). Therefore, we can obtain posterior samples of the survival function by plugging in the posterior samples of $\boldsymbol{\theta}$ in (4.1). Evaluation of the posterior expectation of $S(y \mid \boldsymbol{x})$ is also straightforward through approximation with Monte Carlo integral.

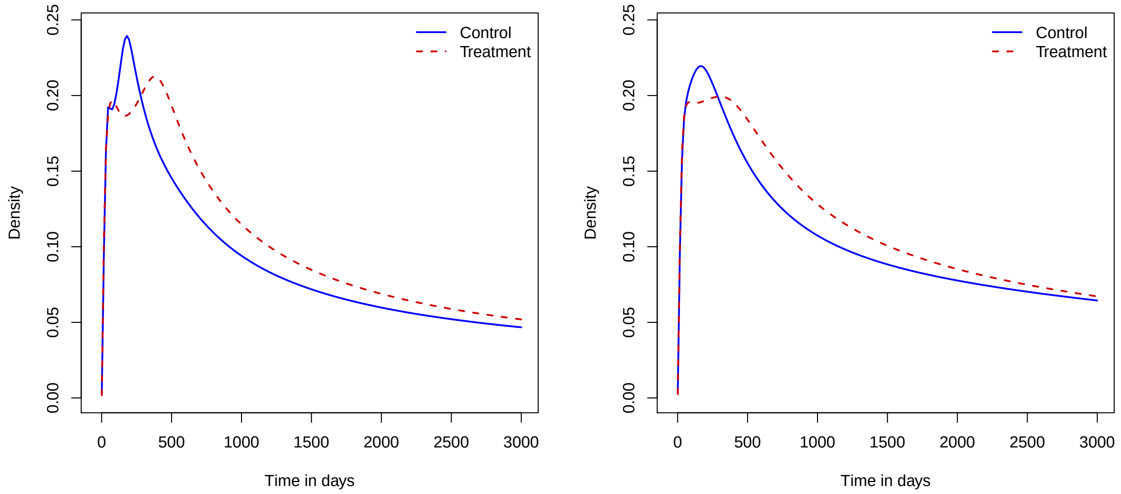
Finally, the inference and implementation for the random-p_k BQMR approach can be readily derived from all above simply by specifying $\gamma_k = 0$ for all k. The same priors for the random-p_k BQMR model for completely observed data apply here. For details, please refer to Section 3.4.1.

4.1.3 Data example: Length of stay at nursing home

We apply the modified BQMR framework for censored data to the nursing home data set analyzed by Morris et al. (1994). Collected from an experiment sponsored by the National Center for Health Services Research in 1980-82, the data set consists of $n = 1601$ observations on the duration of stay (days) of senior patients aged 65 or above at 36 forprofit nursing homes in San Diego, California. Half of those nursing homes were randomized to receive financial incentives for accepting more disabled Medicaid patients as well as for improving a patient's health status and discharging within 90 days, which we consider as "treatment" in this case; while the other half served as the control arm. The data set contains 322 censored observations, which includes 117 on the treatment arm ($n_1 = 712$, censoring rate 16.4%) and 205 on the control arm ($n_0 = 889$, censoring rate 23.1%). We perform an arm-specific identification of influential covariates under the framework and include the following variables as potential predictors for the duration of stay: age, gender, marital status, health condition at admission (binary variable with 1 indicating worse health defined by at least 5 dependencies in activities of daily living) and additional care (required for patients with complications). All covariates are standardized for fair comparison. Since ten subjects in the original data set were admitted and discharged on the same day and therefore had zero length of stay (LOS), to ensure validity of the AFT model, we follow Morris et al. (1994) to add a small constant to LOS and model $y = \log(LOS + 2)$.

We fit each arm with a five-component BQMR framework and consider both the fixed-p_k and random-p_k approach. In terms of prior specification, we use a Laplace prior for $\boldsymbol{\beta}$ with Gamma(0.1,0.1) as hyperprior for η^2, an inverse-Gamma(2,2) prior for σ, truncated normal priors with $\sigma_\mu^2 = 10$ for intercepts μ_{p_k} and a Dirichlet(1,. . . ,1) for the weights $\{\omega_k\}$. For the random-p_k approach, we assume that $\{p_k\}$ follows a homogeneous Poison process a priori. As for the fixed-p_k model, we set $\{p_k\} = \{0.10, 0.25, 0.50, 0.75, 0.90\}$ and apply independent rescaled-Beta priors with shape parameters $\alpha_\gamma = \beta_\gamma = 3$ for γ_k to indicate a slight preference for values that are not too close to the boundaries L_k or U_k.

The posterior inferences are summarized for an "average" patient, defined as having median values for all covariates, which corresponds to an 83-year old female patient, single or widowed, with relatively better health status at admission (less than five dependencies in activity of daily living and not requiring additional care).



(a) Random-p_k BQMR

(b) Fixed-p_k BQMR

Figure 4.1: Nursing home data: Posterior predictive density for length of stay for an average patient under random-p_k BQMR and fixed-p_k BQMR analysis

We plot the posterior predictive density of LOS for such patient under both models in Figure 4.1. There are clearly two modes in the survival time on treatment arm in the random-p_k model. The posterior mean p_k of the last component is 0.51 and 0.54 for the treatment and the control respectively (Figure 4.2). This suggests that the left-to-center of the density function weighs in more in the analysis. Although smoother, the posterior predictive density under the fixed-p_k model delivers a similar message.

The bi-modality in the density function of the treatment cohort gives rise to the intersection of the arm-specific posterior mean survival function and a change of sign in the difference of mean survival function of the two cohorts, as is shown in Figure 4.3 (results

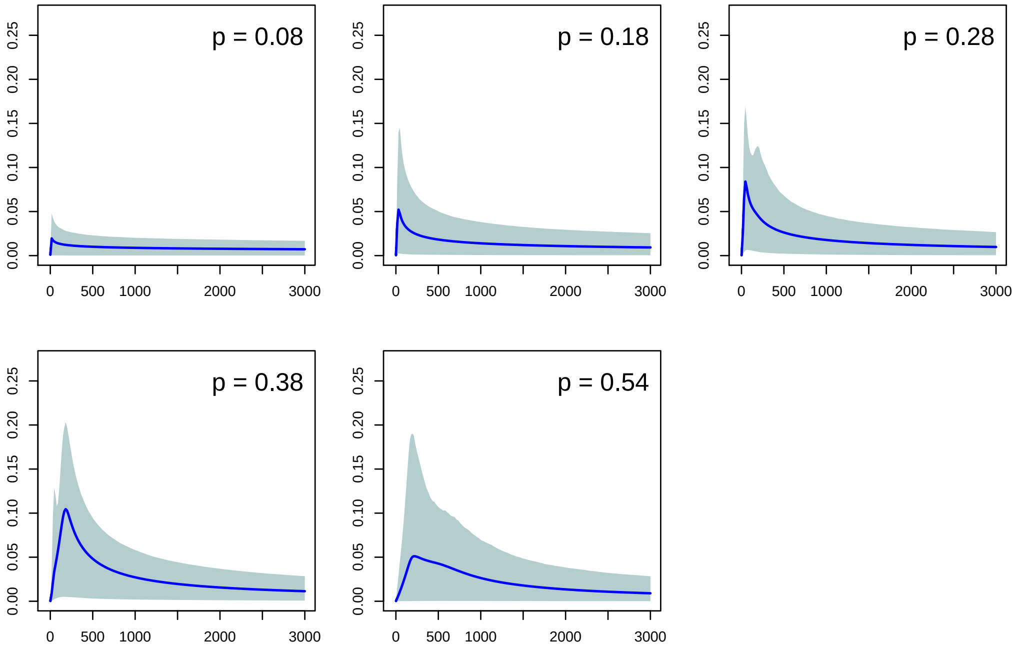
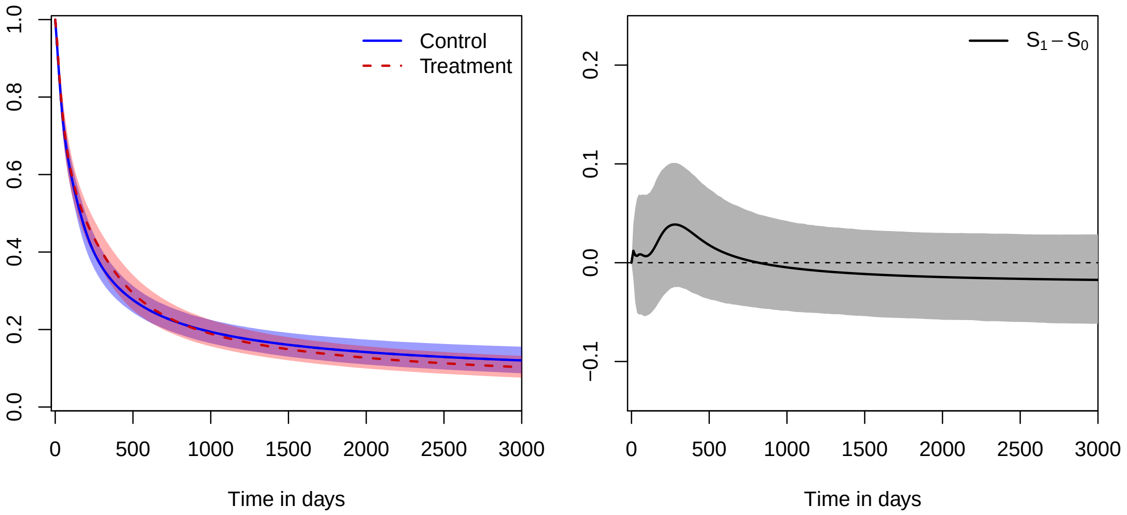


Figure 4.2: Nursing home data: Random-p_k BQMR analysis for control group. Contribution to the density of duration of stay by weighted components (from top left to bottom right: component 1 through 5), mean and 95% CrI. Legend: posterior mean of p_k.

from fixed-p_k is very similar thus omitted). The result suggests that compared with the control, the treatment (financial incentives) did not achieve a substantial improvement of patient health (indicated by reduced LOS). The analysis by Morris et al. (1994) also reached the same conclusion.

Posterior distributions of the median, 75th and 85th percentiles of the length of stay distribution under the two groups support the conclusion that no substantial distinction exists between the treatment and the control (Figure 4.4). The random-p_k and fixed-p_k framework produce quite consistent predictions for the two lower quantiles. For the 85th percentile, the random-p_k model produces a fatter right tail for the control arm (the arm with a higher censoring rate at 23.1%) than the fixed-p_k approach. This suggests that the inference for very high quantiles of the survival distribution may be more volatile if the



(a) Survival function by treatment

(b) Difference in survival function

Figure 4.3: Nursing home data: Posterior inference for an average patient, mean and 95% CrI under the random-p_k model

censoring rate is high.

Finally we compare the predictor identification results from each arm. The posterior summary statistics of the regression coefficients under the fixed-p_k approach is presented in Table 4.1. For this analysis, we consider the predictor as selected by the model if its 95% credible interval does not include zero. Therefore, gender and health condition at admission are selected for both the control and the treatment group. Adjusting for all other covariates, being male and having a poorer health at admission tend to shorten the duration of stay on both arms. For both variables, the effect size is larger in the control group. In addition to gender and initial health status, the model also selected age for the controls and marital status for the treatment group. Among the control cohort, older patients tend to have a substantially longer stay than the younger; while for the treatment group, being married is associated with a shorter LOS. Inference for $\boldsymbol{\beta}$ from the random-p_k model highly resembles

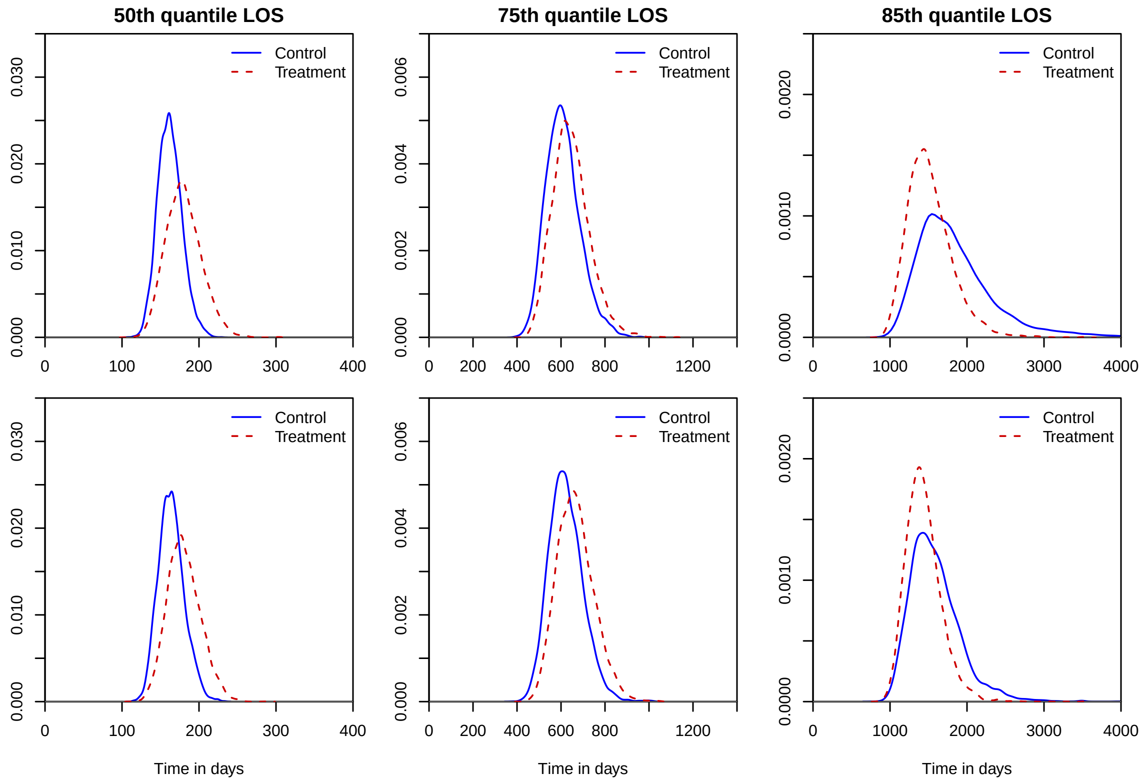


Figure 4.4: Nursing home data: Posterior density of the median, 75th and 85th percentiles of the length of stay distribution for an average patient. Top panel: random-p_k BQMR; bottom panel: fixed-p_k BQMR.

that of its fixed-p_k counterpart.

Variable	Control		Treatment	
Age	**0.133**	(0.006, 0.264)	0.035	(-0.101, 0.173)
Male	**-0.238**	(-0.374, -0.104)	**-0.185**	(-0.325, -0.046)
Married	-0.006	(-0.137, 0.124)	**-0.143**	(-0.285, -0.004)
Poorer health	**-0.290**	(-0.431, -0.147)	**-0.188**	(-0.338, -0.038)
Additional care	-0.080	(-0.209, 0.049)	-0.122	(-0.266, 0.018)

Table 4.1: Nursing home data: Posterior mean and 95% CrI of regression coefficients under fixed-p_k BQMR

Lastly, since the subjects participate in the study at different nursing homes and the randomization is implemented by institution, we recognize a natural hierarchy in the data collection scheme. It would be interesting to include a random effect in the regression analysis to account for the heterogeneity by site. Unfortunately since the data set of this example does not contain the site information, we are unable to carry out the random-effect regression. Nevertheless, the implementation of such model would be quite simple if the site identifiers are available, owing to the fact that both the GAL and the AL kernels are normal mixtures. For sites $i = 1, \ldots, I$, the random effects $b_i \sim N(b_i \mid 0, \sigma_b^2)$ can be estimated with an inverse-gamma prior on σ_b^2 and the posterior full conditional simply follows a Gaussian distribution.

4.2 Modeling disease testing and estimating ROC curve for diagnostic tools

4.2.1 Background

The ROC curve is a graphical plot that measures the diagnostic ability of binary classification system based on a continuous test. It is produced by plotting the true positive rate against the false positive rate at various threshold points for the continuous test result. Consider evaluating a diagnostic tool for a certain disease among the healthy population and the disease population, where $F_0(x)$ denotes the distribution function for the former and $F_1(x)$ represents that of the latter, x being the test score. Let $S_0 = 1 - F_0$ and $S_1 = 1 - F_1$. Then for $u \in [0, 1]$, the ROC curve is defined as,

$$ROC(u) \;=\; S_1\{S_0^{-1}(u)\} \;=\; 1 - F_1\{F_0^{-1}(1-u)\}$$

Further, the area under the curve (AUC) represents the probability that the classifier will generate a higher score to a randomly selected positive (diseased) instance higher than to a randomly chosen negative one (healthy). Defined as the area under the ROC curve, AUC can be calculated as,

$$AUC = \int_0^1 ROC(u)\mathrm{d}u$$

In practice, the distributions for the healthy and the diseased cohort are often quite different. One major reason is that the natural course of many diseases often involves multiple stages and the continuous measure for disease detection (e.g. concentration of antibody) may resemble a step function across stages. Consequently, the test score of the diseased cohort oftentimes has a multi-modal density function. In such case, it is more appropriate to assume two different response distributions for the healthy cohort and the diseased cohort. A flexible modeling scheme then plays an important role in the estimation of ROC curve. Further, in many diseases, the diseased subjects tend to produce higher scores in the diagnostic test than the healthy ones. Thus stochastic ordering may be assumed for the test score distributions of the diseased and the healthy cohorts.

In this section we introduce the application of BQMR for ROC estimation and propose a two-cohort BQMR framework that allows the two cohorts to have different response distribution and its own structure for identifying influential predictors. Additionally, we show that with some careful construction of the model and appropriate specification of the priors, we are able to achieve stochastic ordering in the two-cohort BQMR framework when the data set involves only the test scores and no covariates. For this work, we consider only the gold-standard setting where the disease status is assumed known for all subjects.

4.2.2 Inference for ROC estimation with covariates

When the data set involves predictors, we consider modeling the two cohorts with separate BQMR models, under which both the inference and estimation are then straightforward. Consider n_0 test scores $\boldsymbol{y}_0 = (y_{01}, \ldots, y_{0n_0})$ in the healthy cohort, each associated with a vector of covariates $\boldsymbol{x}_{0i}^T$; and similarly $(\boldsymbol{y}_{1i}, \boldsymbol{x}_{1i}^T)$ for $i = 1, \ldots, n_1$ as the response and predictors of the diseased cohort. It is reasonable to assume two different vectors of regression coefficients, $\boldsymbol{\beta}_0$ and $\boldsymbol{\beta}_1$ for the two cohorts. Denote $\boldsymbol{\theta}_0$ and $\boldsymbol{\theta}_1$ as all the parameter of the BQMR framework of the healthy and the diseased cohort, then the model can be represented as follows,

$$\begin{aligned} y_{0i} \mid \boldsymbol{\beta}_0, \boldsymbol{\theta}_0 &\sim \text{BQMR}(y_{0i} \mid \boldsymbol{x}_{0i}^T \boldsymbol{\beta}_0, \boldsymbol{\theta}_0) \qquad i = 1, \ldots, n_0 \\ y_{1i} \mid \boldsymbol{\beta}_1, \boldsymbol{\theta}_1 &\sim \text{BQMR}(y_{1i} \mid \boldsymbol{x}_{1i}^T \boldsymbol{\beta}_1, \boldsymbol{\theta}_1) \qquad i = 1, \ldots, n_1 \end{aligned}$$

where $BQMR(\eta, \boldsymbol{\theta})$ represents either a fixed-p_k or a random-p_k BQMR framework with regression function η and parameters $\boldsymbol{\theta}$.

Each cohort can be fitted separately with MCMC algorithms. Using M posterior samples of $\boldsymbol{\theta}_0$ and $\boldsymbol{\theta}_1$, we can easily obtain M samples of survival functions S_0 and S_1 (see equation (4.1)) and numerically evaluate S_0^{-1}. Consider $\boldsymbol{u}_{l=i,\ldots,L}$ as n threshold points over the unit interval. The mth posterior samples of the ROC curve is simply

$$ROC^{(m)}(u_l) = S_1^{(m)}\{S_0^{-1(m)}(u_l)\}, \qquad l = 1, \ldots, L$$

Then the area under the curve can be approximated with a Riemann sum,

$$AUC^{(m)} = \sum_{l=0}^{L-1} ROC^{(m)}(u_l)(u_{l+1} - u_l)$$

where $u_0 = 0$.

4.2.3 Inference for ROC estimation without covariates

The purpose of this section is to build a BQMR framework with stochastic ordering for the ROC application. We focus on the scenario where the data involve only the response variable and no covariates as a simpler application to develop the method.

When the data set consists of solely the test scores, it is possible to construct a two-cohort BQMR model satisfying the assumption of stochastic ordering. We start by deriving some theoretical results firstly on the stochastic ordering of GAL distributions, then on that of BQMR framework. The definition of stochastic ordering is the following: Let X and Y be two random variables with distributions on the real line. If $P(X > x) \leq P(Y > x)$ for all $x \in \mathbb{R}$, then X is said to be *smaller than Y in the usual stochastic order* (denoted by $X \leq_{st} Y$) (Shaked & Shanthikumar, 2007).

Theoretical results on stochastic ordering for BQMR

The following lemmas provide theoretical results on the stochastic ordering of GAL distribution, AL distribution and the mixture distributions and serve as the foundation of the two-cohort BQMR model to be proposed.

Lemma 1. Stochastic ordering of $\mathbf{GAL}_{p_0}$ distributions by location μ. Let Y_1 follow a $GAL_{p_0}(\mu_1, \gamma, \sigma)$ distribution and Y_2 follow a $GAL_{p_0}(\mu_2, \gamma, \sigma)$ distribution for any fixed p_0, $0 < p_0 < 1$. If $\mu_1 \leq \mu_2$, then $Y_1 \leq_{st} Y_2$.

Proof. Denote F_1 and F_2 as the distribution function for Y_1 and Y_2, respectively. Under the hierarchical representation of GAL distributions, we can write

$$\begin{aligned}
F_1(y) &= \int_{-\infty}^{y}\int_{\mathbb{R}^+}\int_{\mathbb{R}^+} N(t \mid \mu_1 + \sigma C|\gamma|s + \sigma Az, \sigma^2 Bz) Exp(z|1) N^+(s|0,1)\mathrm{d}z\,\mathrm{d}s\,\mathrm{d}t \\
&= \int_{\mathbb{R}^+}\int_{\mathbb{R}^+}\int_{-\infty}^{y} N(t \mid \mu_1 + \sigma C|\gamma|s + \sigma Az, \sigma^2 Bz)\mathrm{d}t\, Exp(z|1)\mathrm{d}z\, N^+(s|0,1)\mathrm{d}s \\
&= \int_{\mathbb{R}^+}\int_{\mathbb{R}^+} \Phi(y \mid \mu_1 + \sigma C|\gamma|s + \sigma Az, \sigma^2 Bz) Exp(z|1)\mathrm{d}z\, N^+(s|0,1)\mathrm{d}s \\
F_2(y) &= \int_{\mathbb{R}^+}\int_{\mathbb{R}^+} \Phi(y \mid \mu_2 + \sigma C|\gamma|s + \sigma Az, \sigma^2 Bz) Exp(z|1)\mathrm{d}z\, N^+(s|0,1)\mathrm{d}s
\end{aligned}$$

Denote $\Phi_k(y \mid z,s) = \Phi(y \mid \mu_k + \sigma C|\gamma|s + \sigma Az, \sigma^2 Bz)$ for $k = 1$ and 2. $\forall z > 0,\ s > 0$, if $\mu_1 \leq \mu_2$, then $\Phi_1(y \mid z,s) \geq \Phi_2(y \mid z,s)$ by the stochastic ordering of normal distributions by location. Then the following holds for marginalizing z,

$$\begin{aligned}
G_1(y \mid s) &= \int_{\mathbb{R}^+} \Phi_1(y \mid z,s) Exp(z \mid 1)\mathrm{d}z \\
&= \int_{\mathbb{R}^+} \Phi_2(y \mid z,s) Exp(z \mid 1)\mathrm{d}z + \int_{\mathbb{R}^+} [\Phi_1(y \mid z,s) - \Phi_2(y \mid z,s)] Exp(z \mid 1)\mathrm{d}z \\
&\geq \int_{\mathbb{R}^+} \Phi_2(y \mid z,s) Exp(z \mid 1)\mathrm{d}z \;=\; G_2(y \mid s)
\end{aligned}$$

Finally by integrating out s, we get

$$\begin{aligned}
F_1(y) &= \int_{\mathbb{R}^+} G_1(y \mid s) N^+(s|0,1)\mathrm{d}s \\
&= \int_{\mathbb{R}^+} G_2(y \mid s) N^+(s|0,1)\mathrm{d}s + \int_{\mathbb{R}^+} [G_1(y \mid s) - G_2(y \mid s)] N^+(s|0,1)\mathrm{d}s \\
&\geq \int_{\mathbb{R}^+} G_2(y \mid s) N^+(s|0,1)\mathrm{d}s \;=\; F_2(y)
\end{aligned}$$

By definition of stochastic ordering, $Y_1 \leq_{st} Y_2$.

Lemma 2. Stochastic ordering of AL$_p$ distributions by skewness p. Let Y_1 follow an $AL_{p_1}(\mu,\sigma)$ distribution and Y_2 follow a $AL_{p_2}(\mu,\sigma)$ distribution. If $0 < p_2 \leq p_1 < 1$, then $Y_1 \leq_{st} Y_2$.

Proof. The distribution function of AL distribution is given by,

$$F(y) \quad = \quad \begin{cases} p \exp\left\{\dfrac{1-p}{\sigma}(y-\mu)\right\}, & y-\mu \leq 0 \\ 1-(1-p)\exp\left\{-\dfrac{p}{\sigma}(y-\mu)\right\}, & y-\mu > 0 \end{cases}$$

Denote F_1 and F_2 as the distribution function for Y_1 and Y_2, respectively. Let S_1 and S_2 be the survival function of Y_1 and Y_2. Given that $0 < p_2 \leq p_1 < 1$, depending on the sign of $y - \mu$, there are the following two cases:

i) If $y - \mu \leq 0$, then

$$\frac{F_2(y)}{F_1(y)} \quad = \quad \frac{p_2 \exp\left\{\frac{1-p_2}{\sigma}(y-\mu)\right\}}{p_1 \exp\left\{\frac{1-p_1}{\sigma}(y-\mu)\right\}} \quad = \quad \frac{p_2}{p_1} \exp\left\{\frac{p_1-p_2}{\sigma}(y-\mu)\right\} \quad \leq \quad \frac{p_2}{p_1} \quad \leq \quad 1$$

ii) If $y - \mu > 0$, then

$$\begin{aligned} \frac{S_2(y)}{S_1(y)} &= \frac{(1-p_2)\exp\left\{-\frac{p_2}{\sigma}(y-\mu)\right\}}{(1-p_1)\exp\left\{-\frac{p_1}{\sigma}(y-\mu)\right\}} = \frac{1-p_2}{1-p_1}\exp\left\{-\frac{p_2-p_1}{\sigma}(y-\mu)\right\} \\ &\geq \frac{1-p_2}{1-p_1} \geq 1 \\ \Rightarrow \frac{F_2(y)}{F_1(y)} &= \frac{1-S_2(y)}{1-S_1(y)} \leq 1 \end{aligned}$$

Thus we have shown that $F_1(y) \geq F_2(y)$. By definition of stochastic ordering, $Y_1 \leq_{st} Y_2$.

Remark. Lemma 1 extends the analogous result for the Gaussian distribution to the GAL family, the former being the most common example of a parametric family stochastically ordered by its location. Focusing on two AL distributions with the same location parameter, Lemma 2 is not of practical significance to our modeling. However, we report it as an independent result which we use to prove the next lemma.

Lemma 3. Stochastic ordering of $\mathbf{AL}_p$ distributions by skewness p and location μ. Let Y_1 follow an $AL_{p_1}(\mu_1, \sigma)$ distribution and Y_2 follow a $AL_{p_2}(\mu_2, \sigma)$ distribution. If $0 < p_2 \leq p_1 < 1$ and $\mu_1 \leq \mu_2$, then $Y_1 \leq_{st} Y_2$.

Proof. Since AL distribution belongs to the GAL family, lemma 1 on the stochastic ordering of GAL_{p_0} by location μ also applies to the AL distributions. Denote F_1 and F_2 as the distribution function for Y_1 and Y_2 and let F^{AL} represent the distribution function of an AL random variable. Given that $\mu_1 \leq \mu_2$ and $0 < p_2 \leq p_1 < 1$, we can show that

$$\begin{aligned}
F_1(y) \;=\; F^{AL}(y \mid p_1, \mu_1, \sigma) \;&\geq\; F^{AL}(y \mid p_1, \mu_2, \sigma) \qquad \text{(lemma 1 : stochastic ordering by location)} \\
&\geq\; F^{AL}(y \mid p_2, \mu_2, \sigma) \qquad \text{(lemma 2 : stochastic ordering by skewness)} \\
&=\; F_2(y)
\end{aligned}$$

By definition, $Y_1 \leq_{st} Y_2$. Therefore, we have established the stochastic ordering of AL distributions if the location and scale are ordered in opposite directions.

Lemma 4. Stochastic ordering of mixture distributions. Consider two sets of real-valued random variables $\{Y_{11}, Y_{12}, \ldots, Y_{1K}\}$ with density functions $\{f_{11}, f_{12}, \ldots, f_{1K}\}$ and $\{Y_{21}, Y_{22}, \ldots, Y_{2K}\}$ with density functions $\{f_{21}, f_{22}, \ldots, f_{2K}\}$, $K \in \mathbb{N}$. Let Y_1 follow a mixture distribution of the form $\sum_{k=1}^{K} \omega_k f_{1k}(y_1)$ and Y_2 follow a mixture distribution of the form $\sum_{k=1}^{K} \omega_k f_{2k}(y_2)$, where weights $\sum_{k=1}^{K} \omega_k = 1$. If for $k = 1, \ldots, K$ there is $Y_{1k} \leq_{st} Y_{2k}$, namely if there exists pairwise stochastic ordering between $\{Y_{1k}\}$ and $\{Y_{2k}\}$, then $Y_1 \leq_{st} Y_2$.

Proof. Denote $\{F_{11}, F_{12}, \ldots, F_{1K}\}$ as the distribution functions of $\{Y_{11}, Y_{12}, \ldots, Y_{1K}\}$ and denote $\{F_{21}, F_{22}, \ldots, F_{2K}\}$ for those of $\{Y_{21}, Y_{22}, \ldots, Y_{2K}\}$. Since $Y_{1k} \leq_{st} Y_{2k}$ for $k = 1, \ldots, K$, by definition of stochastic ordering, $\forall y \in \mathbb{R}$ there is

$$F_1(y) \;=\; \omega_1 F_{11}(y) + \omega_2 F_{12}(y) + \ldots + \omega_K F_{1K}(y)$$

$$\begin{aligned} &\geq \omega_1 F_{21}(y) + \omega_2 F_{22}(y) + \ldots + \omega_K F_{2K}(y) \qquad (F_{1k}(y) \geq F_{2k}(y) \text{ for } k = 1, \ldots, K) \\ &= F_2(y) \end{aligned}$$

By definition of stochastic ordering, $Y_1 \leq_{st} Y_2$.

We apply the above results to the fixed-p_k and random-p_k BQMR framework and deduce the following lemma, which is the last lemma in this section and also the key theoretical result on stochastic ordering for BQMR models.

Lemma 5. Stochastic ordering of mixture of GAL and AL distributions. Consider random variables Y_1 and Y_2 from two K-component mixture distributions, $K \in \mathbb{N}$.

i. **GAL kernel.** Define Y_1 and Y_2 as mixtures of GAL distributions, such that $f(y_1) = \sum_{k=1}^{K} \omega_k f_{p_0}^{GAL}(y_1 \mid \mu_1, \sigma)$ and $f(y_2) = \sum_{k=1}^{K} \omega_k f_{p_0}^{GAL}(y_2 \mid \mu_2, \sigma)$. If $\mu_{1k} \leq \mu_{2k}$ for $k = 1, \ldots, K$, then $Y_1 \leq_{st} Y_2$.

ii. **AL kernel.** Define Y_1 and Y_2 as mixtures of AL distributions, such that $f(y_1) = \sum_{k=1}^{K} \omega_k f_{p_1}^{AL}(y_1 \mid \mu_1, \sigma)$ and $f(y_2) = \sum_{k=1}^{K} \omega_k f_{p_2}^{AL}(y_2 \mid \mu_2, \sigma)$. If $p_{2k} \leq p_{1k}$ and $\mu_{1k} \leq \mu_{2k}$ for $k = 1, \ldots, K$, then $Y_1 \leq_{st} Y_2$.

Proof. Result i follows by applying Lemma 1 and Lemma 4. Result ii follows by applying Lemma 3 and Lemma 4.

Lemma 5 shows that we can achieve stochastic ordering on random variables following mixture of GAL or AL distributions if we carefully construct the mixture and impose order constraints on the location (and the skewness) of the components. The result on the AL distribution is more flexible than that of the GAL, allowing both μ_k and p_k to vary between the cohorts for given k. Therefore, we propose a two-cohort framework with stochastic ordering based on the random-p_k BQMR in the next section.

The stochastic ordering results of the BQMR framework we obtain is based upon the pairwise ordering of the quantile components. An alternative common approach to achieve stochastic ordering of mixture distributions is to construct two stochastically ordered weight vectors (see Theorem 1.A.6 in Shaked & Shanthikumar (2007)). We did not follow that path in this work, because the AL and the GAL distributions, as kernel distribution, do not satisfy the monotonicity prerequisite of the theorem on the joint parameter space of percentage p_0 and location μ. Instead of constructing structured weights, we apply the same weight configuration to both cohorts and explore the stochastic ordering property of the kernels. We will show in Section 4.2.5 with a real data example that when the two cohorts have dramatically different response distributions, the proposed framework with a common weight vector for both cohorts can still achieve satisfactory goodness-of-fit, which in a way demonstrates the flexibility of the model.

BQMR model with random-p_k and stochastic ordering

When the diagnostic test data does not involve any predictors, we propose to approach the estimation of ROC by modeling the test score distribution of the healthy and the diseased cohorts in a joint framework. Consider n_0 observations, $(y_{0i}, \ldots, y_{0n_0})$, in the healthy cohort; and n_1 observations, $(y_{0i}, \ldots, y_{0n_1})$, in the diseased cohort. We fit the data from each cohort with a K-component random-p_k BQMR model and construct priors that depend on each other, assuming common weight allocation $\{w_k\}$ and scale σ, but different $\{p_k\}$ and $\{\mu_{p_k}\}$. We impose order constraints $\mu_{p_{0k}} \leq \mu_{p_{1k}}$ and $p_{0k} \geq p_{1k}$, $k = 1, \ldots, K$. It follows from Lemma 5 that under such framework, the test scores of the healthy cohort is smaller than that of the diseased in the usual stochastic order.

We introduce latent indicator variables ξ_{0ik}, $i = 1, \ldots, n_0$, $k = 1, \ldots, K$ and ξ_{1ik}, $i = 1, \ldots, n_1$, $k = 1, \ldots, K$ for each cohort. For ease of notation, we denote $\boldsymbol{\theta}_0 = (\{p_{0k}\}, \{\mu_{p_{0k}}\})$ and $\boldsymbol{\theta}_1 = (\{p_{1k}\}, \{\mu_{p_{1k}}\})$ and set $\mu_{p_{00}} = \mu_{p_{10}} = -\infty$ and $\mu_{p_{0(K+1)}} = \mu_{p_{1(K+1)}} = \infty$. The model in its mixture representation can be expressed as,

$$\begin{aligned}
y_{0i} \mid \boldsymbol{\theta}_0, \sigma, v_{0i}, \boldsymbol{\xi}_{0i} &\sim \prod_{k=1}^{K} \left[N(y_{0i} \mid \mu_{p_{0k}} + A_{p_{0k}} v_i, \sigma B_{p_{0k}} v_i)\right]^{\xi_{0ik}} & i = 1, \ldots, n_0 \\
y_{1i} \mid \boldsymbol{\theta}_1, \sigma, v_{1i}, \boldsymbol{\xi}_{1i} &\sim \prod_{k=1}^{K} \left[N(y_{1i} \mid \mu_{p_{1k}} + A_{p_{1k}} v_i, \sigma B_{p_{1k}} v_i)\right]^{\xi_{1ik}} & i = 1, \ldots, n_1 \\
\boldsymbol{\xi}_{0i} \mid \omega_1, \ldots, \omega_k &\sim Multinomial(\boldsymbol{\xi}_{0i} \mid \omega_1, \ldots \omega_k) & i = 1, \ldots, n_0 \\
\boldsymbol{\xi}_{1i} \mid \omega_1, \ldots, \omega_k &\sim Multinomial(\boldsymbol{\xi}_{1i} \mid \omega_1, \ldots \omega_k) & i = 1, \ldots, n_1 \\
\boldsymbol{v}_0, \boldsymbol{v}_1 \mid \sigma &\sim \prod_{i=1}^{n_0} Exp(v_{0i} \mid 1/\sigma) \prod_{i=1}^{n_1} Exp(v_{1i} \mid 1/\sigma)
\end{aligned}$$

For the intercepts and the percentages of each cohort, we take the priors in Chapter 3, then add the restrictions involving the corresponding component in the other cohort. The priors for the joint framework are as follows,

$$\begin{aligned}
\pi(\{\mu_{p_{0k}}\}, \{\mu_{p_{1k}}\}) &\propto \prod_{k=1}^{K} N(\mu_{p_{0k}} \mid \mu_{0k}, \sigma_\mu^2) \mathbb{1}\{\mu_{p_{0(k-1)}} < \mu_{p_{0k}} < \mu_{p_{0(k+1)}}\} \\
&\quad \prod_{k=1}^{K} N(\mu_{p_{1k}} \mid \mu_{1k}, \sigma_\mu^2) \mathbb{1}\{\mu_{p_{1(k-1)}} < \mu_{p_{1k}} < \mu_{p_{1(k+1)}}\} \prod_{k=1}^{K} \mathbb{1}\{\mu_{p_{0k}} \leq \mu_{p_{1k}}\} \\
\pi(\{p_{0k}\}, \{p_{1k}\}) &\propto \mathbb{1}\{0 < p_{01} < \ldots < p_{0k} < \ldots < p_{0K} < 1\} \\
&\quad \mathbb{1}\{0 < p_{11} < \ldots < p_{1k} < \ldots < p_{1K} < 1\} \prod_{k=1}^{K} \mathbb{1}\{p_{0k} \geq p_{1k}\} \\
\pi(\omega_1, \ldots, \omega_K) &\propto Dir(a_1, \ldots, a_K)
\end{aligned}$$

The posterior samples can be obtained via MCMC sampling. The sampling algorithm for most parameters is essentially the same for the one cohort model, except for the location and the skewness parameters, which can be sampled with the following steps,

1. Sample $\mu_{p_{0k}}$ from $N(\mu^*_{p_{0k}}, (\sigma^*_{p_{0k}})^2)\mathbb{1}\{\mu_{p_{0(k-1)}} < \mu_{p_{0k}} < \min\{\mu_{p_{0(k+1)}}, \mu_{p_{1k}}\}\}$, for $k = 1, \ldots, K$; then sequentially sample $\mu_{p_{1k}}$ from $N(\mu^*_{p_{1k}}, (\sigma^*_{p_{1k}})^2)\mathbb{1}\{\max\{\mu_{p_{1(k-1)}}, \mu_{p_{0k}}\} < \mu_{p_{1k}} < \mu_{p_{1(k+1)}}\}$, where for $d = 0$ and $d = 1$ there is

$$(\sigma^*_{p_{dk}})^2 \;=\; \left[\frac{1}{\sigma^2_\mu} + \sum_{\xi_{dik}=1} \frac{1}{B_{di}\sigma v_{di}}\right]^{-1}, \;\; \mu^*_{p_{dk}} \;=\; (\sigma^*_{p_{dk}})^2 \sum_{\xi_{dik}=1} \frac{y_{di} - A_{di}v_{di}}{B_{di}\sigma v_{di}}$$

2. Sample p_{0k} and p_{1k} with Metropolis-Hastings steps. Denote $p_0 = 0$ and $p_{K+1} = 1$, then the posterior full conditional of p_{0k} is proportional to

$$\prod_{i=1}^{n_d} \left[N(y_{0i} \mid \mu_{p_{0k}} + A_{p_{0k}}v_{0i}, \sigma B_{p_{0k}}v_{0i})\right]^{\xi_{0ik}} \mathbb{1}\{\max\{p_{0(k-1)}, p_{1k}\} < p_{0k} < p_{0(k+1)}\}$$

and that of p_{1k} is proportional to

$$\prod_{i=1}^{n_1} \left[N(y_{1i} \mid \mu_{p_{1k}} + A_{p_{1k}}v_{1i}, \sigma B_{p_{1k}}v_{1i})\right]^{\xi_{1ik}} \mathbb{1}\{p_{1(k-1)} < p_{1k} < \min\{p_{1(k+1)}, p_{0k}\}\}$$

4.2.4 Data example: Adolescent depression (with covariates)

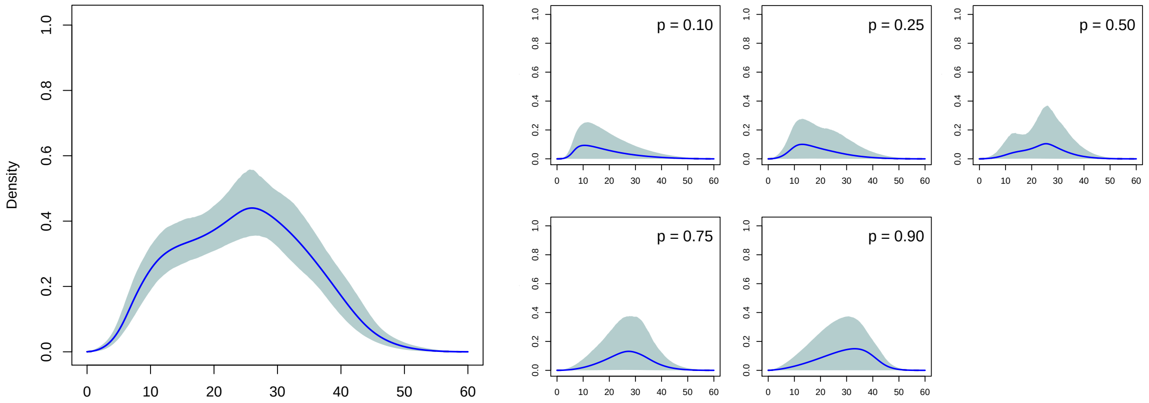
We illustrate the method with a data set from a depression study in 1986 (Addy et al., 1994) conducted to evaluate the diagnosis of adolescent depression with the Center for Epidemiologic Studies Depression Scale (CES-D), which is a 20-item self-report rating scale widely used to measure depression symptomatology in the adult population. The data set contains information from $n = 458$ seventh and eighth graders, including depression diagnosis (gold standard), CES-D score (ranges between 0 and 60), cohesion score (a measure of emotional bonding to the family ranging from 16 to 80) and demographic information.

The diseased cohort consists of $n_1 = 72$ youths and the remaining $n_0 = 386$ depression-free subjects constitute the healthy cohort.

We include gender (female/male), race (white/black) and cohesion score (standardized with mean within the cohort and a common scale of 10 points for ease of interpretation) as predictors and model the CES-D score on the logit-scale as response by each cohort. With interests in both tails and the center of the response, we fit each cohort with a five-component fixed-p_k BQMR framework, with $\{p_k\} = \{0.1, 0.25, 0.5, 0.75, 0.9\}$. We used the same priors as in the survival example in Section 4.1.3 for both cohorts.

We summarize the posterior inferences in Figure 4.5 and 4.6 for an "average" patient with median covariate value for all covariates, which translates to a white female with a cohesion score of 51. The obvious difference in the posterior predictive densities for the healthy and the diseased demonstrates the need to fit the two cohorts separately: the predictive conditional distribution of the CES-D score is much flatter in the healthy cohort, while the diseased has a more concentrated predictive distribution with substantial left-skewness. The by-component contribution confirms that the tails and the center are almost equally important for the healthy cohort. For the diseased, the center of the response distribution receives much heavier weight than the tails.

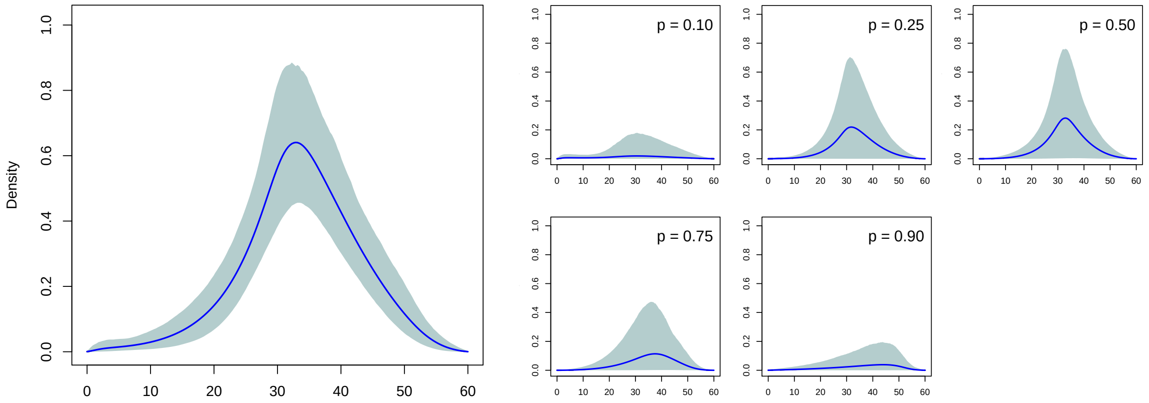
The cumulative distribution function by cohort and the ROC curve are shown in Figure 4.7. The CES-D separates the two distributions quite well over most of its range, implying promising diagnostic application of the instrument among adolescent depression patients. This is confirmed by a mean AUC estimate of 0.753 with 95% credible interval (0.676, 0.819).



(a) Predictive density

(b) Contribution by weighted components

Figure 4.5: Depression data: Fixed-p_k BQMR analysis. Posterior inference on the CES-D score of an average patient without depression, mean and 95% CrI.

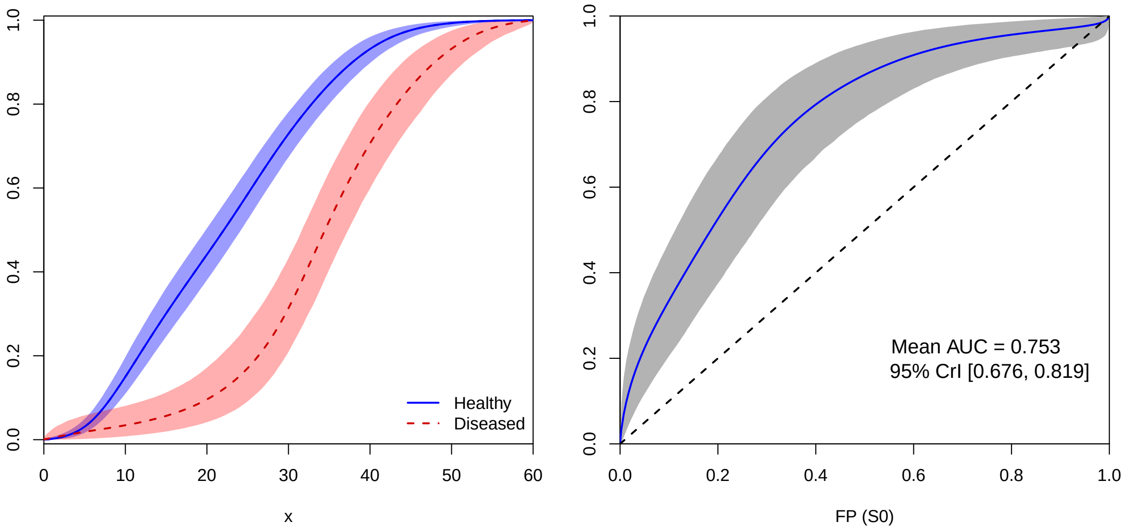


(a) Predictive density

(b) Contribution by weighted components

Figure 4.6: Depression data: Fixed-p_k BQMR analysis. Posterior inference on the CES-D score of an average patient with depression, mean and 95% CrI.

A summary of the posterior inference for regression coefficients of each cohort is presented in Table 4.2. In both cohorts, gender exhibits a substantial effect on predicting the CES-D score with the 95% CrI clearly away from zero, while race does not seem to be affecting the response much. In the healthy cohort, increasing cohesion score is associated with lower CES-D score. However, there is insufficient evidence to support the same relationship for the depression subjects. The result suggests a differential effect of the cohesion



(a) Cumulative distribution function by cohort

(b) ROC curve

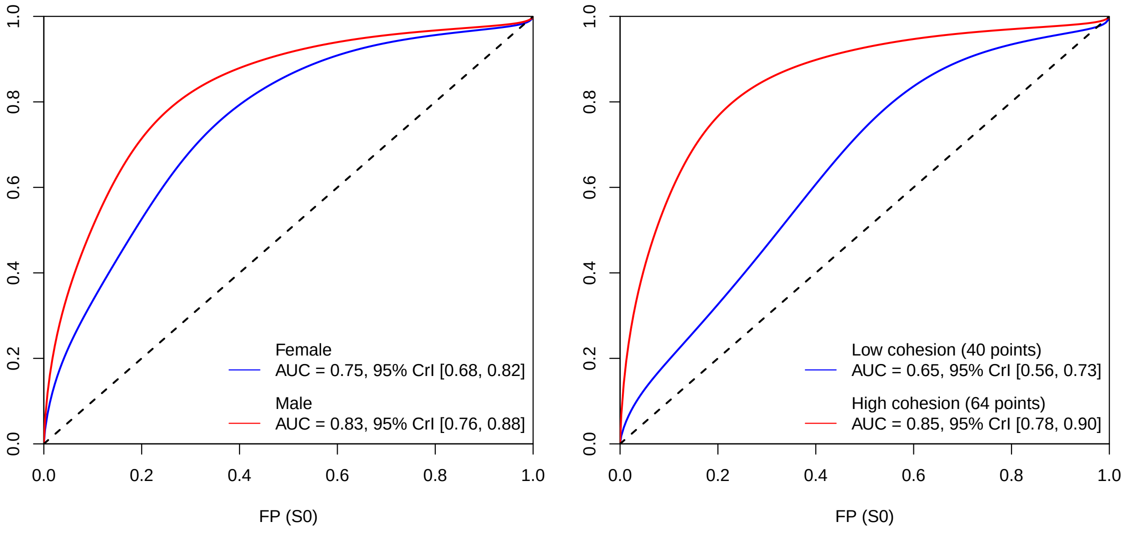
Figure 4.7: Depression data: Fixed-p_k BQMR analysis. Posterior inference for the CDF of CES-D score by cohort (left) and the ROC and AUC (right) of an average patient, mean and 95% CrI

score by cohort on the prediction of CES-D.

Variable	Healthy		Diseased	
Male	**-0.318**	(-0.500, -0.128)	**-0.790**	(-1.178, -0.315)
Race (black)	-0.047	(-0.261, 0.171)	0.220	(-0.147, 0.647)
Cohesion (unit: 10 pts)	**-0.444**	(-0.552, -0.335)	-0.044	(-0.233, 0.146)

Table 4.2: Depression data: Posterior mean and 95% CrI of regression coefficients under fixed-p_k BQMR

Finally, we compare the covariate-adjusted ROC estimates by gender and by cohesion score in Figure 4.8. For an average patient (white with cohesion score of 51), the CES-D measure appears to be a quite good diagnostic tool for adolescent depression regardless of gender, with a higher AUC when applied among young men. It preserves a very good accuracy when the average subject (white female) has a high cohesion score, or a strong emotional bonding with her family. However, when the cohesion score is low, the



(a) Posterior mean ROC by gender

(b) Posterior mean ROC by cohesion score

Figure 4.8: Depression data: Posterior inference for ROC and AUC with different inputs in gender (left) and in cohesion (right) for an average patient under fixed-p_k BQMR.

diagnostic capability of CES-D weakens.

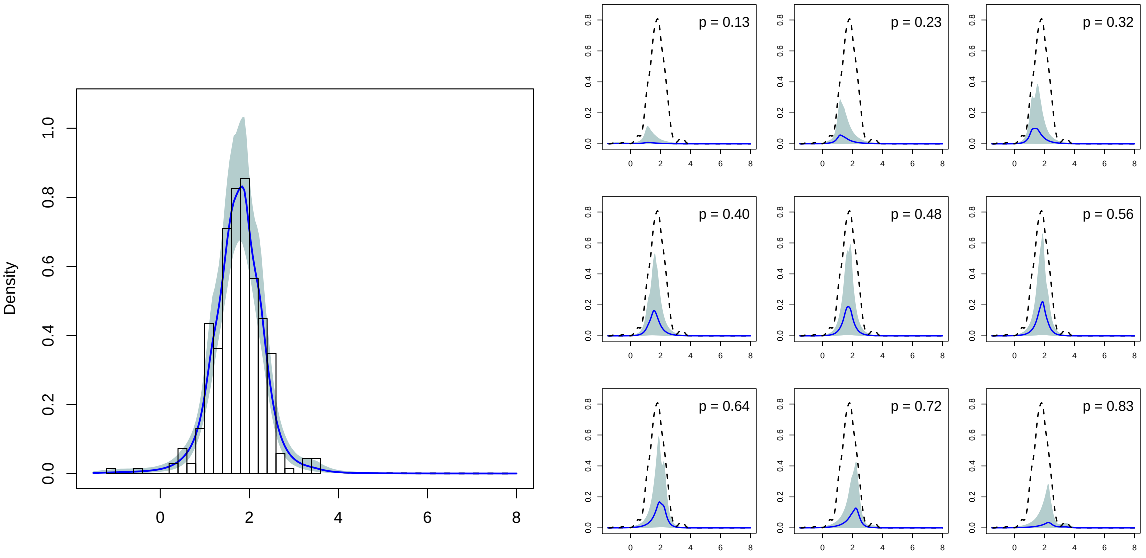
4.2.5 Data example: Johnes disease (without covariates)

Johnes disease (Mycobacterium avium paratuberculosis (MAP)) is a contagious fatal disease that mainly affects the small intestines of ruminants. Accurate diagnosis of the disease plays an important role in the successful surveillance and effective control of the symptoms and the spread of the epidemic. In particular, evaluation of antibody in the blood sample with ELISA kits is one of the most popular detection technologies of MAP.

We analyze a serologic data set from $n_0 = 345$ disease-free cattle (labeled as "healthy") and $n_1 = 258$ diseased cattle studied in Hanson et al. (2008) to model the distribution of antibody by disease status and estimate the ROC curve. We begin with fitting each cohort with a separate BQMR framework, assuming independence between the two. Then we fit all observations together with the two-cohort BQMR model proposed in the previous section and impose stochastic ordering on the serologic score.

Method 1: Independent BQMR models for each cohort

Both the random-p_k framework and the fixed-p_k approach are applied in this analysis. In both cases, we consider $K = 9$ components, with $p_k = k/10$, $k = 1, \ldots, 9$ for the fixed-p_k BQMR. Common priors are used for both cohorts, including an inverse-gamma(2,2) for scale σ, a Dirichlet$(1, \ldots, 1)$ prior for the weights and truncated normal priors with mean 0 and $\sigma_\mu^2 = 10$ for intercepts μ_k. For the fixed-p_k model, independent Beta(3,3) priors are applied for γ_k on the logit scale.

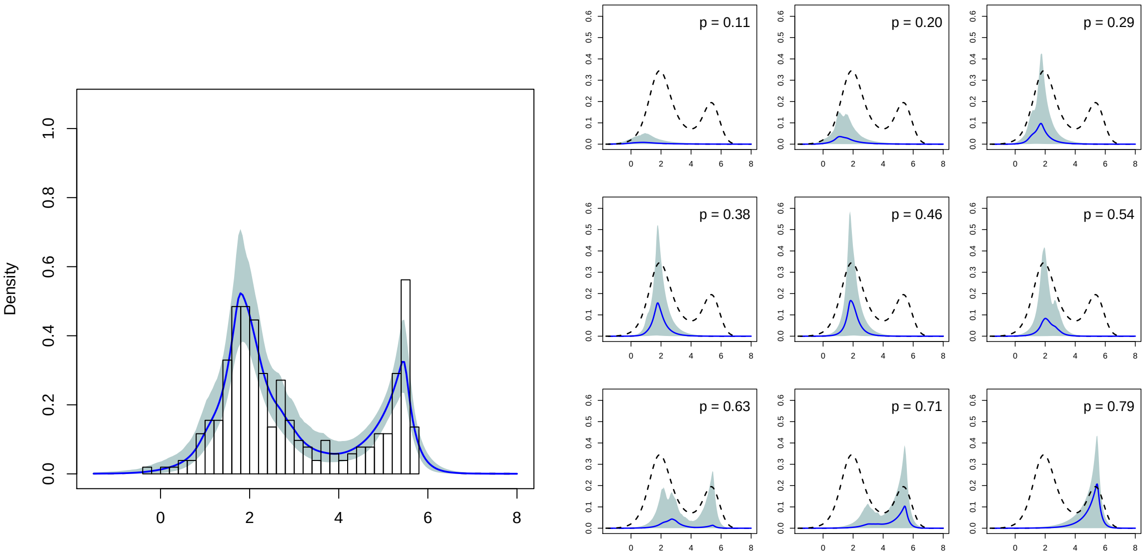


(a) Posterior density overlayed with data

(b) Contribution from weighted components

Figure 4.9: MAP data: Random-p_k BQMR analysis for the healthy cohort. Posterior density of ELISA reading overlayed with histogram of the data (left) and the contribution from each weighted component (right), mean and pointwise 95% interval

Figure 4.9 and 4.10 present the predictive densities of the serologic score for each cohort from the random-p_k framework, as well as the contribution from each weighted components. The model captures the behavior in the data quite well for both cohorts and there is a clear difference in the posterior predictive density between the healthy and the diseased cattle. The predictive scores are unimodal for the healthy cohort. While for the cattle with MAP, we observe two modes in the serologic reading, which is likely caused by the difference in antibody level associated with different stages of the disease.

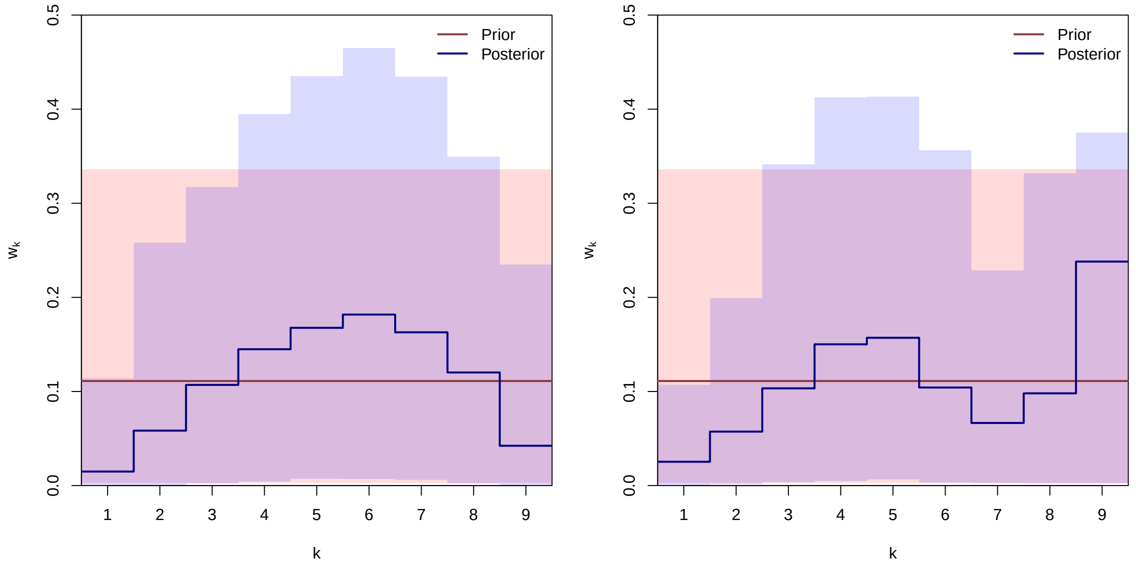


(a) Posterior density overlayed with data

(b) Contribution from weighted components

Figure 4.10: MAP data: Random-p_k BQMR analysis for the diseased cohort. Posterior density of ELISA reading overlayed with histogram of the data (left) and the contribution from each weighted component (right), mean and pointwise 95% interval

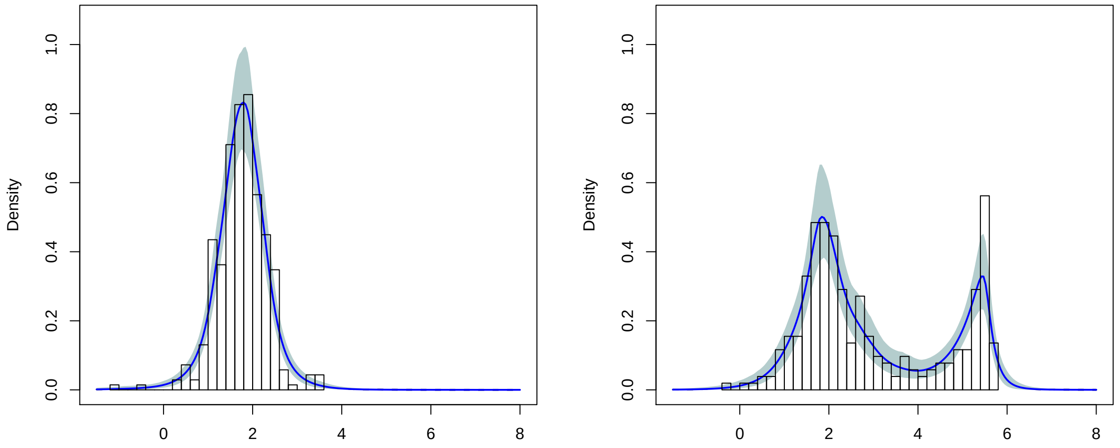
Figure 4.11 shows the posterior inference of $\{w_k\}$ overlayed with the prior under the random-p_k BQMR. The solid blue line plots the posterior mean of w_k by component k associated with the 95% interval marked with the light blue bands. With a heavy weight in the 9th component, the model shows more emphasis on the right tail of the serological score for the diseased cohort, which contrasts the model for the disease-free cattle with more weight on the center of the response distribution. Results from the fixed-p_k approach are show in Figure 4.12 and 4.13 and are in general consistent with the random-p_k model.



(a) Healthy cohort

(b) Diseased cohort

Figure 4.11: MAP data: Prior and posterior mean and 95% CrI of w_k by component of the health cohort (left) and the diseased cohort (right) under random-p_k BQMR.

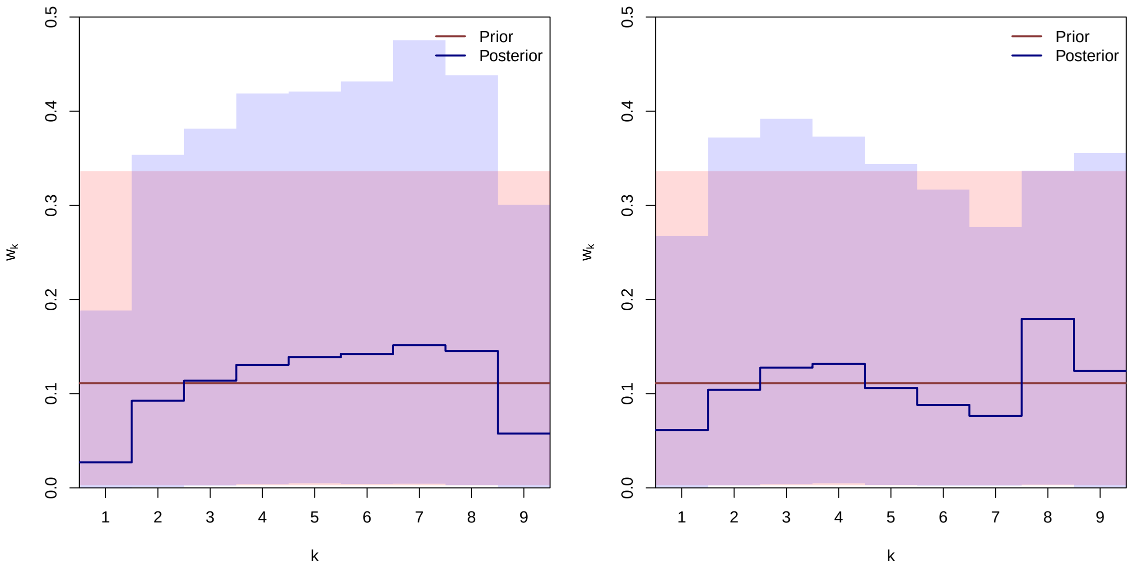


(a) Healthy cohort

(b) Diseased cohort

Figure 4.12: MAP data: Posterior mean and pointwise 95% interval of posterior densities of ELISA reading overlayed with histogram of the data by cohort under fixed-p_k BQMR.

We estimate the ROC curve and the cumulative distribution function (CDF) of the serologic ELISA test using the posterior samples from each model (Figure 4.14 and 4.15). The posterior mean and 95% credible interval of the area under the curve (AUC) are



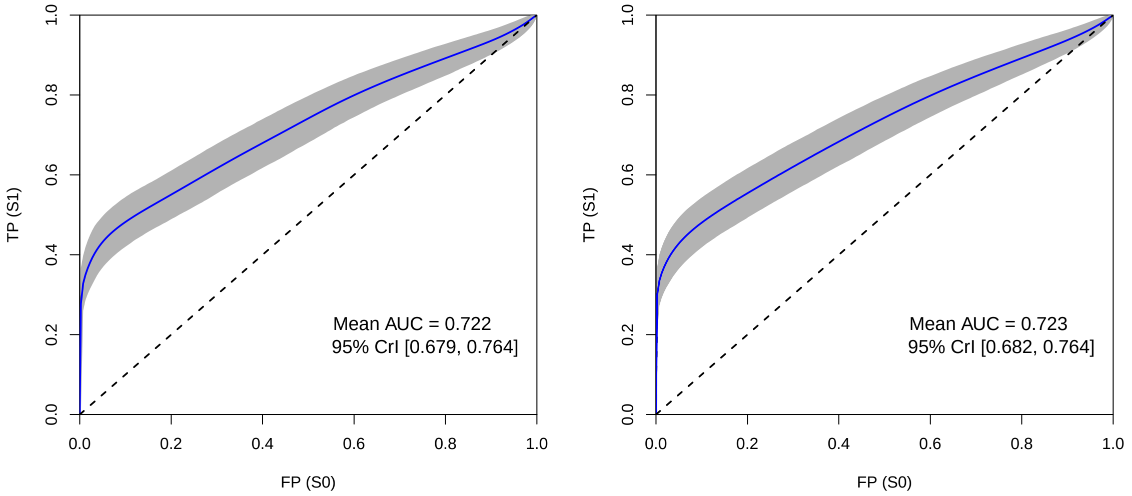
(a) Healthy cohort

(b) Diseased cohort

Figure 4.13: MAP data: Prior and posterior mean and 95% CrI of w_k by component of the health cohort (left) and the diseased cohort (right) under fixed-p_k BQMR.

also presented with the ROC plots. The random-p_k model and the fixed-p_k approach agree quite well on the estimation of AUC, with the posterior mean being roughly 0.72 under both models.

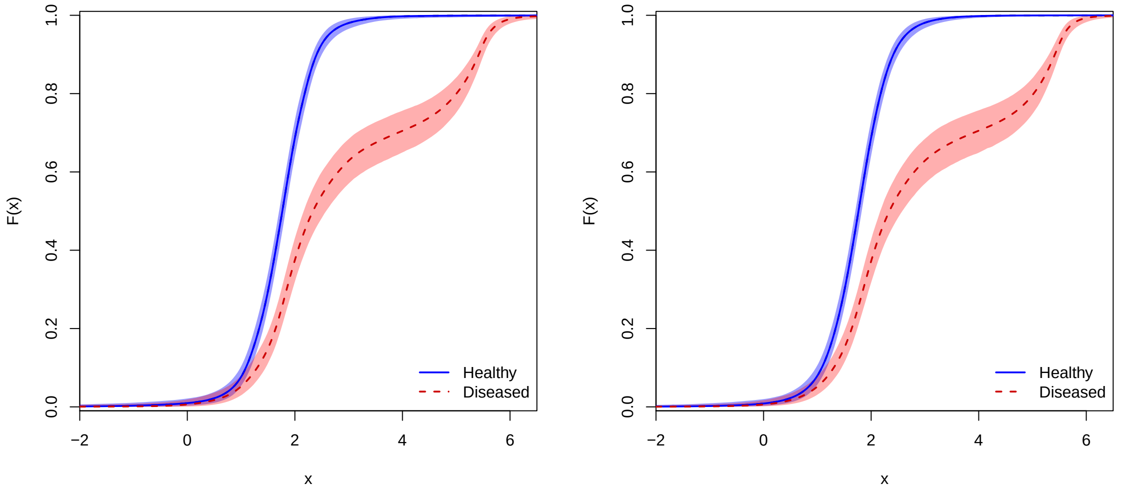
We compute the log-pseudo-marginal-likelihood (LPML) (Geisser & Eddy, 1979; Gelfand et al., 1992; Gelfand & Dey, 1994) to assess the model fitting (Table 4.3). Larger LPML values represent better model-fitting from a predictive standpoint. Overall the goodness-of-fit is comparable between the two methods, with the fixed-p_k approach attaining a slightly higher LPML for the diseased cohort and the combined measure than the random-p_k model.



(a) Random-p_k BQMR

(b) Fixed-p_k BQMR

Figure 4.14: MAP data: Posterior mean and 95% CrI of ROC and AUC by model. No assumptions on stochastic ordering.



(a) Random-p_k BQMR

(b) Fixed-p_k BQMR

Figure 4.15: MAP data: Posterior mean and 95% CrI of CDF of ELISA reading of each cohort by model. No assumptions on stochastic ordering.

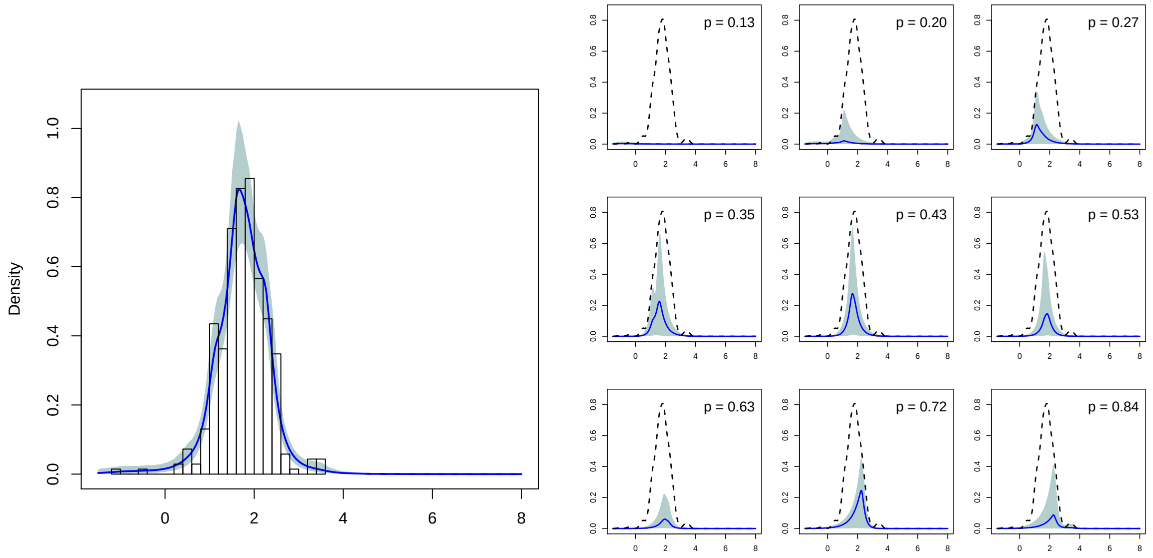
Model	$LPML_0$	$LPML_1$	$\sum LPML$	*AUC*
Random-p_k BQMR, $K = 9$	-272.0	-391.6	-663.6	0.722 (0.679,0.764)
Fixed-p_k BQMR, $p_k = k/10$, $k = 1, \ldots, 9$	-272.4	-388.9	-661.3	0.723 (0.682,0.764)

Table 4.3: MAP data: Log-pseudo-marginal-likelihood for the health cohort (LPML_0), the diseased cohort (LPML_1 and combined ($\sum$LPML) and mean AUC with 95% CrI by model. No assumptions on stochastic ordering.

Method 2: Two-cohort BQMR models with stochastic ordering

Now we apply the two-cohort random-p_k model and impose the ordering in distribution, such that the response (serologic reading) of the healthy cohort is stochastically smaller than that of the diseased. Wherever applicable, same priors are used as in the previous analysis of independent fitting.

The posterior predictive density of the serologic score shows a decent fit from the two-cohort approach (Figure 4.16 and 4.17). Since now the two cohorts share the same weights, the posterior mean of w_k is substantially higher in both the center and the right tail than in the other areas (Figure 4.18), which is likely a consequence of applying common weights to accommodate the behavior of the response in both cohorts.

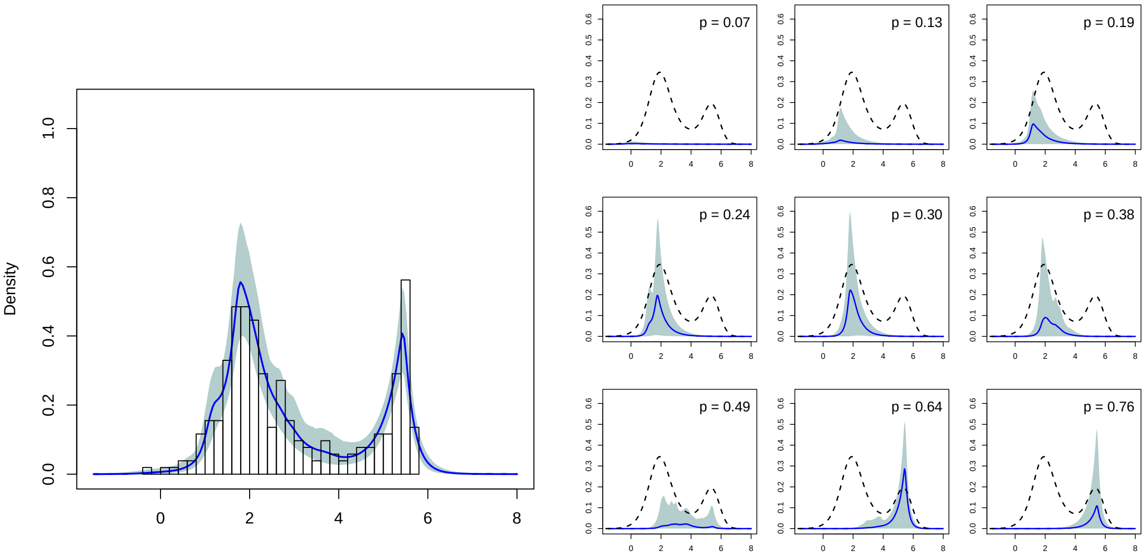


(a) Predictive density overlayed with data

(b) Contribution from weighted components

Figure 4.16: MAP data: Two-cohort BQMR analysis for the healthy cohort. Posterior density of ELISA reading overlayed with histogram of the data (left) and the contribution from each weighted component (right), mean and pointwise 95% interval.

From both the ROC curve and CDF of the two cohorts in Figure 4.19, we see that



(a) Predictive density overlayed with data

(b) Contribution from weighted components

Figure 4.17: MAP data: Two-cohort BQMR analysis for the diseased cohort. Posterior density of ELISA reading overlayed with histogram of the data (left) and the contribution from each weighted component (right), mean and pointwise 95% interval.

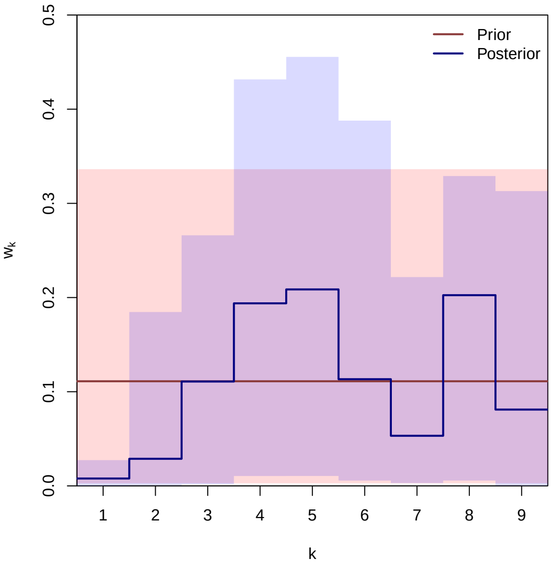
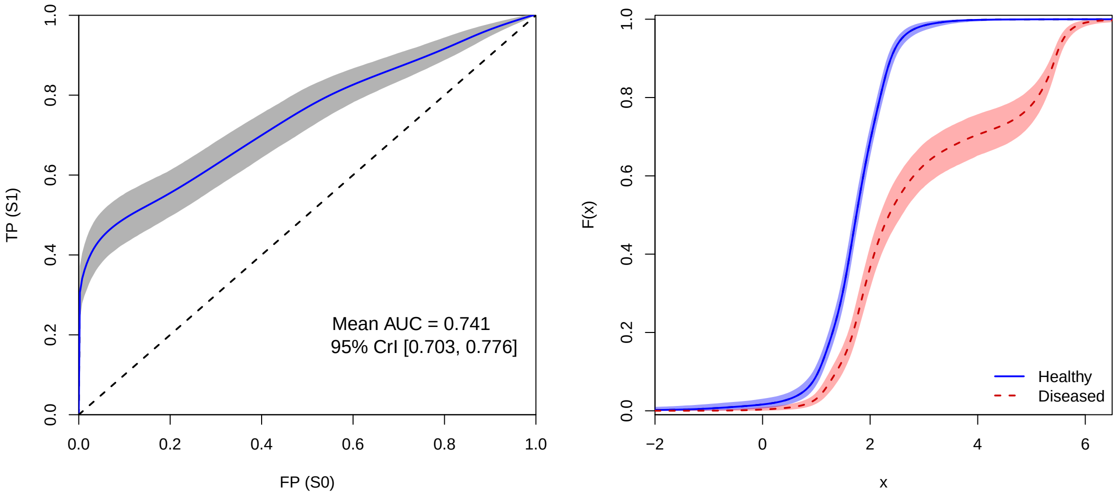


Figure 4.18: MAP data: Prior and posterior mean and 95% CrI of w_k by component under two-cohort BQMR.

the serologic score for the healthy cattle is stochastically smaller than that of the diseased. The posterior mean of AUC is 0.741, which is higher than the result in the previous analysis. We also notice that the point-wise 95% interval for the ROC curve is slightly narrower than in the earlier result where the cohorts are fitted separately. The same applies to the 95% interval of the AUC.



(a) ROC

(b) CDF by cohort

Figure 4.19: MAP data: Two-cohort BQMR analysis with stochastic ordering assumption. Posterior mean and 95% CrI of ROC and AUC (left) and CDF of ELISA reading of each cohort (right).

Finally, the LPML of the two-cohort approach indicates a decent fit on the data (Table 4.4). If we compare it with the previous results, we can see that the LPML values are actually very close to what we get by fitting the cohorts separately. This suggests that the proposed two-cohort BQMR framework achieves stochastic ordering in the response variable with minimal compromise in the goodness-of-fit of this data. In other word, stochastic ordering of the diseased test score larger than that of the healthy cattle appears to be a reasonable assumption for the analysis of the MAP data.

Model	*LPML*$_0$	*LPML*$_1$	$\sum$*LPML*	*AUC*
Two-cohort random-p_k BQMR, $K = 9$	-272.1	-390.6	-662.7	0.741 (0.703,0.776)

Table 4.4: MAP data: Two-cohort BQMR analysis with stochastic ordering assumption. Log-pseudo-marginal-likelihood for the health cohort (LPML$_0$), the diseased cohort (LPML$_1$ and combined ($\sum$LPML) and mean AUC with 95% CrI.

4.3 Discussion

In this chapter we explored extensions of the BQMR framework with applications in survival analysis and ROC curve estimation. Through fitting the two cohorts with separate BQMR models, we achieve arm-specific identification of important variables taking into account the influence of covariates on multiple parts of the response distribution. Moreover, we derive theoretical results on stochastic ordering for the BQMR models. In a no-covariate scenario, enlightened by the theoretical findings, we construct the stochastically ordered two-cohort BQMR framework to model the test score of both the diseased and the disease-free cohort in a coherent manner. Evaluation of the posterior predictive performance of approaches with and without assumptions on stochastic ordering offers a way to check the assumptions we make for the two cohorts.

When modeling the survival responses, we augment the data set with latent observations and carry out inferences on the augmented parameter space. Another possible approach is to work with the survival functions directly. However, posterior sampling under this method requires implementations of Metropolis-Hastings steps for almost all parameters, which will cause very poor mixing of the Markov chain. Although the data augmentation approach we take involves posterior sampling of a good number of latent variables, we benefit from the nice convergence property of Gibbs sampling. With a decent number of observations and a reasonable censoring proportion, we observe quick convergence of the MCMC and efficient estimation of the parameters, as is shown in the nursing data example.

We developed the two-cohort BQMR model with stochastic ordering assumption for data sets without covariates as a direct application of Lemma 5. However, the theoretical results are actually not restricted to the no-covariate scenario. When the two cohorts share exactly the same set of covariates, Lemma 5 will still apply if we can enforce some constraints on the regression coefficients, such that the linear predictors of the two cohorts are ordered given the same covariate values. For instance, if there is only one covariate x and x is bounded from below, we can offset all x_i by a constant to make them non-negative. By restricting the regression coefficients of the two cohorts to be both positive and ordered, we can easily apply Lemma 5 and ensure stochastic ordering between the cohorts. More complicated constraints are needed as the number of covariates increases.

Chapter 5

Conclusions

In this dissertation, we develop a new Bayesian regression framework based on a structured mixture of quantile regressions. Using both simulation study and real data examples, we demonstrate its performance in identifying influential predictors and illustrate its application in approximating the underlying true conditional response distribution. The main contribution and also the highlights of this work is that we approach the identification of important covariates in a Bayesian paradigm through modeling the response distribution with a collection of quantile regressions, from which we synthesize the inferences and obtain a combined estimation of the covariate effects taking into account how the predictors influence multiple parts of the response distribution. The construction of the framework with its focus on quantile regression imparts nice interpretations to the model. Further, the visualization of the contribution to the posterior predictive error density from each quantile regression component offers a graphical way to acquire a better understanding of how different parts of the response distribution comes into play both in the regression analysis and in the predictive inference.

Our work begins with the development of a flexible kernel density for the BQMR framework, which is parameterized in terms of the quantile and is more flexible than the commonly used AL distribution. If viewed from a modeling perspective, it is obvious that as an error distribution, the AL distribution imposes very strong assumptions on the behavior of the density function, due to the fact that it is a single parameter distribution of which the skewness is fully tied to the quantile. Therefore under the assumption of AL errors, it is challenging to obtain reasonable posterior predictive inference regarding the conditional response distribution. The work in Chapter 2 is intended for addressing this problem. We introduce a shape parameter through modifying the hierarchical representation of the AL distribution and manage to link the location of the distribution to the quantile of interest. Owing to the extra parameter, for any given probability p, the resulting GAL distribution can have varying skewness, mode and tail behavior as the shape parameter γ changes, making it more flexible than the AL counterpart. The latter turns out to be a special case of the GAL family. We offer the GAL distribution as an alternative to the AL distribution for Bayesian quantile regression, focusing on quantile estimation and predictive inference for the response distribution at the same time, which we demonstrate through both simulation studies and real data examples.

We would like to also emphasize some key features that distinguishes the GAL distribution from other potentially more flexible approaches, such as Bayesian nonparametric methods. As a parametric distribution developed from a normal mixture, the GAL distribution has a very straightforward MCMC sampling scheme. It can be easily adapted for modeling various types of responses, such as censored data in the Tobit quantile regression extension we present in Section 2.2.3. Moreover, compared with the Bayesian nonparametric methods, the relatively simple parametric structure of the GAL distribution makes it a decent candidate for the building block of some more complicated modeling framework. In fact, the development of BQMR in the Chapter 3 and its extensions and applications in Chapter 4 are good examples along the line.

The idea of BQMR was originally motivated by the composite quantile regression in the classical literature (Zou & Yuan, 2008), although eventually the BQMR framework is developed in a similar spirit but constructed with a very different formulation. In Zou & Yuan (2008), the CQR estimator under adaptive lasso penalty is shown to be oracular and exhibits good performance in variable selection, but as a pure optimization procedure it is difficult to transplant the composite regression to a modeling framework. Zhao et al. (2016) attempted to translate the CQR estimation directly into a Bayesian paradigm, the result being that the method they develop uses the data multiple times and therefore not a valid probabilistic model. Retaining the idea of a common set of regression coefficients shared between different quantile regressions, we navigate slightly away from matching the loss function of CQR and instead resort to an additive mixture of multiple quantile regression components with structured priors. The advantage of taking this path is that with a valid and highly flexible conditional distribution for the response variable, the framework ensures that the estimation and inference are all conducted under a well-defined probabilistic setting. We have developed two versions of BQMR, one with the GAL kernels and the other with AL components. The GAL framework is intended for combined estimation of covariates when a given set of $\{p_k\}$ is of interest, and the AL model applies to the complementary situation. With the $\{p_k\}$ varying, the latter also explores the response distribution in a more efficient manner.

To make the BQMR framework more flexible, we allow the weights $\{w_k\}$ to vary and further generalize the fixed-p_k GAL model to take a nonparametric prior for the weights. In real life applications, it can be quite the opposite that the researcher may be interested in a set of $\{p_k\}$ with certain weight configuration. Informative priors can be applied to the weight vector under these circumstances. In the AL version of the model, we also place a non-informative HPP prior on $\{p_k\}$ for flexibility consideration. A potential drawback of this prior is that the first and last percentage in $\{p_k\}$ can get arbitrarily close to the two end points, 0 and 1, with nonzero probability a priori, which may inflate the variability the posterior inference. A more informative prior, such as a non-homogeneous Poisson process, will be a better choice in the case where certain $\{p_k\}$ are favored a priori.

In Chapter 4 we explore applications of the proposed BQMR model in survival analysis and evaluation of a binary classifier/diagnostic tool in biomedical sciences. The data examples of both applications show that a good number of posterior inferences can be drawn under the proposed BQMR model, such as for the survival function at different quantiles and for covariate-dependent ROC estimation. Moreover, we are able to incorporate stochastic ordering in a two-cohort BQMR model for ROC estimation without covariates. It offers a way to evaluate the stochastic ordering assumption between two cohorts in covariate-free ROC studies. For instance, the leave-one-out goodness-of-fit metric LPML confirms good model fitting of the approach in the data example in section 4.2.5.

The fact that both the GAL and AL distributions are normal mixtures makes the BQMR framework highly amenable, thanks to the nice properties of the Gaussian distribution. In this dissertation, we focus on modeling the integrated covariate effects with linear regression. In fact, in a scenario where the predictor effect is highly nonlinear, it is possible to associate the BQMR framework with the idea of a Gaussian process. The two could connect nicely under the unravelled the hierarchical representation of the quantile regression kernels. Future work can be directed to the exploration of such generalization.

Bibliography

Addy, C., Jackson, K., McKeown, R., Waller, J. & Garrison, C. (1994). Two stage sampling designs for adolescent depression studies.

Amemiya, T. (1984). Tobit models: A survey. *Journal of econometrics* **24**, 3–61.

Azzalini, A. (1985). A class of distributions which includes the normal ones. *Scandinavian journal of statistics* pp. 171–178.

Cade, B. S. & Noon, B. R. (2003). A gentle introduction to quantile regression for ecologists. *Frontiers in Ecology and the Environment* **1**, 412–420.

Chib, S. (1992). Bayes inference in the tobit censored regression model. *Journal of Econometrics* **51**, 79–99.

Das, P. & Ghosal, S. (2017). Bayesian quantile regression using random b-spline series prior. *Computational Statistics & Data Analysis* **109**, 121–143.

Embrechts, P., Klüppelberg, C. & Mikosch, T. (1997). *Modelling extremal events*, volume 33. Springer Science & Business Media.

FERGUSON, T. S. (1973). A bayesian analysis of some nonparametric problems. *The annals of statistics* pp. 209–230.

FERNÁNDEZ, C. & STEEL, M. F. (1998). On bayesian modeling of fat tails and skewness. *Journal of the American Statistical Association* **93**, 359–371.

GEISSER, S. & EDDY, W. F. (1979). A predictive approach to model selection. *Journal of the American Statistical Association* **74**, 153–160.

GELFAND, A. E. & DEY, D. K. (1994). Bayesian model choice: asymptotics and exact calculations. *Journal of the Royal Statistical Society. Series B (Methodological)* pp. 501–514.

GELFAND, A. E., DEY, D. K. & CHANG, H. (1992). Model determination using predictive distributions with implementation via sampling-based methods. Technical report, STANFORD UNIV CA DEPT OF STATISTICS.

GELFAND, A. E. & GHOSH, S. K. (1998). Model choice: a minimum posterior predictive loss approach. *Biometrika* **85**, 1–11.

GERACI, M. & BOTTAI, M. (2007). Quantile regression for longitudinal data using the asymmetric Laplace distribution. *Biostatistics* **8**, 140–154.

GREEN, P. J. (1995). Reversible jump markov chain monte carlo computation and bayesian model determination. *Biometrika* **82**, 711–732.

HANSON, T. E., KOTTAS, A. & BRANSCUM, A. J. (2008). Modelling stochastic order in the analysis of receiver operating characteristic data: Bayesian non-parametric approaches. *Journal of the Royal Statistical Society: Series C (Applied Statistics)* **57**, 207–225.

Harrison, D. & Rubinfeld, D. L. (1978). Hedonic housing prices and the demand for clean air. *Journal of environmental economics and management* **5**, 81–102.

Henze, N. (1986). A probabilistic representation of the'skew-normal'distribution. *Scandinavian journal of statistics* pp. 271–275.

Hjort, N. L. & Petrone, S. (2007). Nonparametric quantile inference using Dirichlet processes. *Advances in statistical modeling and inference* pp. 463–492.

Hjort, N. L. & Walker, S. G. (2009). Quantile pyramids for Bayesian nonparametrics. *The Annals of Statistics* pp. 105–131.

Hoti, F. & Sillanpää, M. (2006). Bayesian mapping of genotype× expression interactions in quantitative and qualitative traits. *Heredity* **97**, 4–18.

Hu, Y., Zhao, K. & Lian, H. (2013). Bayesian quantile regression for partially linear additive models. *Statistics and Computing* pp. 1–18.

Huang, H. & Chen, Z. (2015). Bayesian composite quantile regression. *Journal of Statistical Computation and Simulation* **85**, 3744–3754.

Koenker, R. (2005). *Quantile Regression.* New York: Cambridge University Press.

Koenker, R. & Bassett Jr, G. (1978). Regression quantiles. *Econometrica: journal of the Econometric Society* pp. 33–50.

KOTTAS, A. & GELFAND, A. E. (2001). Bayesian semiparametric median regression modeling. *Journal of the American Statistical Association* **96**, 1458–1468.

KOTTAS, A. & KRNJAJIĆ, M. (2009). Bayesian semiparametric modelling in quantile regression. *Scandinavian Journal of Statistics* **36**, 297–319.

KOZUMI, H. & KOBAYASHI, G. (2011). Gibbs sampling methods for Bayesian quantile regression. *Journal of statistical computation and simulation* **81**, 1565–1578.

LEE, D. & NEOCLEOUS, T. (2010). Bayesian quantile regression for count data with application to environmental epidemiology. *Journal of the Royal Statistical Society: Series C (Applied Statistics)* **59**, 905–920.

LI, Q., XI, R., LIN, N. et al. (2010). Bayesian regularized quantile regression. *Bayesian Analysis* **5**, 533–556.

LUM, K., GELFAND, A. E. et al. (2012). Spatial quantile multiple regression using the asymmetric Laplace process. *Bayesian Analysis* **7**, 235–258.

MORRIS, C., NORTON, E. & ZHOU, X. (1994). Parametric duration analysis of nursing home usage. *Case Studies in Biometry* pp. 231–248.

MROZ, T. A. (1987). The sensitivity of an empirical model of married womenś hours of work to economic and statistical assumptions. *Econometrica: Journal of the Econometric Society* pp. 765–799.

NOUFAILY, A. & JONES, M. (2013). Parametric quantile regression based on the generalized gamma distribution. *Journal of the Royal Statistical Society: Series C (Applied Statistics)* **62**, 723–740.

Park, T. & Casella, G. (2008). The bayesian lasso. *Journal of the American Statistical Association* **103**, 681–686.

Petrone, S. (1999). Bayesian density estimation using bernstein polynomials. *Canadian Journal of Statistics* **27**, 105–126.

Prentice, R. L., Kalbfleisch, J. D., Peterson Jr, A. V., Flournoy, N., Farewell, V. & Breslow, N. (1978). The analysis of failure times in the presence of competing risks. *Biometrics* pp. 541–554.

Raftery, A. E. (1996). Approximate bayes factors and accounting for model uncertainty in generalised linear models. *Biometrika* **83**, 251–266.

Reich, B. J. (2012). Spatiotemporal quantile regression for detecting distributional changes in environmental processes. *Journal of the Royal Statistical Society: Series C (Applied Statistics)* **61**, 535–553.

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

Reich, B. J. & Smith, L. B. (2013). Bayesian quantile regression for censored data. *Biometrics* **69**, 651–660.

Scaccia, L. & Green, P. J. (2003). Bayesian growth curves using normal mixtures with nonparametric weights. *Journal of Computational and Graphical Statistics* **12**, 308–331.

SCOTT, S. L. (2002). Bayesian methods for hidden markov models: Recursive computing in the 21st century. *Journal of the American Statistical Association* **97**, 337–351.

SHAKED, M. & SHANTHIKUMAR, G. (2007). *Stochastic orders*. Springer Science & Business Media.

SHERLOCK, C., FEARNHEAD, P. & ROBERTS, G. O. (2010). "the random walk metropolis: Linking theory and practice through a case study". *Statistical Science* pp. 172–190.

SRIRAM, K., RAMAMOORTHI, R., GHOSH, P. et al. (2013). Posterior consistency of bayesian quantile regression based on the misspecified asymmetric laplace density. *Bayesian Analysis* **8**, 479–504.

TADDY, M. A. & KOTTAS, A. (2010). A bayesian nonparametric approach to inference for quantile regression. *Journal of Business & Economic Statistics* **28**.

THOMPSON, P., CAI, Y., MOYEED, R., REEVE, D. & STANDER, J. (2010). Bayesian nonparametric quantile regression using splines. *Computational Statistics & Data Analysis* **54**, 1138–1150.

TIBSHIRANI, R. (1996). Regression shrinkage and selection via the lasso. *Journal of the Royal Statistical Society. Series B (Methodological)* pp. 267–288.

TOKDAR, S. & KADANE, J. B. (2011). Simultaneous linear quantile regression: A semiparametric Bayesian approach. *Bayesian Analysis* **6**, 1–22.

TSIONAS, E. G. (2003). Bayesian quantile inference. *Journal of statistical computation and simulation* **73**, 659–674.

Volinsky, C. T. & Raftery, A. E. (2000). Bayesian information criterion for censored survival models. *Biometrics* **56**, 256–262.

Walker, S. G. & Mallick, B. K. (1999). A Bayesian semiparametric accelerated failure time model. *Biometrics* **55**, 477–483.

Wang, M. & Zhang, L. (2012). A Bayesian quantile regression analysis of potential risk factors for violent crimes in usa. *Open Journal of Statistics* **2**, 526.

Wichitaksorn, N., Choy, B. S. T. & Gerlach, R. (2014). A generalized class of skew distributions and associated robust quantile regression models. *The Canadian Journal of Statistics* **42**, 579–596.

Yu, K. & Moyeed, R. A. (2001). Bayesian quantile regression. *Statistics & Probability Letters* **54**, 437–447.

Yu, K. & Stander, J. (2007). Bayesian analysis of a tobit quantile regression model. *Journal of Econometrics* **137**, 260–276.

Yu, K. & Zhang, J. (2005). A three-parameter asymmetric Laplace distribution and its extension. *Communications in StatisticsTheory and Methods* **34**, 1867–1879.

Zhao, W.-h., Zhang, R.-q., Lü, Y.-z. & Liu, J.-c. (2016). Bayesian regularized regression based on composite quantile method. *Acta Mathematicae Applicatae Sinica, English Series* **32**, 495–512.

Zhu, D. & Galbraith, J. W. (2011). Modeling and forecasting expected shortfall with the generalized asymmetric student-t and asymmetric exponential power distributions. *Journal of Empirical Finance* **18**, 765–778.

Zhu, D. & Zinde-Walsh, V. (2009). Properties and estimation of asymmetric exponential power distribution. *Journal of Econometrics* **148**, 86–99.

Zou, H. & Yuan, M. (2008). Composite quantile regression and the oracle model selection theory. *The Annals of Statistics* pp. 1108–1126.

Appendix A

MCMC Algorithms

A.1 Regularized BQR with lasso under GAL distribution

We apply a $Ga(a_{\eta^2}, b_{\eta^2})$ prior for η^2. Then the algorithm is very similar to the penalty-free version provided in Section 2.2.1, except for $\boldsymbol{\beta}$, which will be sampled from a $N(\boldsymbol{\beta}^*,\, \Sigma^*)$, with

$$\Sigma^* \quad = \quad \left(\sum_{i=1}^{n} \frac{\boldsymbol{x_i} \boldsymbol{x_i}^T}{B\sigma v_i} + \Omega \right)^{-1}, \quad \boldsymbol{\beta}^* = \Sigma^* \left\{ \sum_{i=1}^{n} \frac{\boldsymbol{x}_i [y_i - (\beta_0 + \sigma C|\gamma|\, s_i + Av_i)]}{B\sigma v_i} \right\}$$

where $\Omega = \mathrm{diag}(\omega_1^{-1}, \ldots, \omega_d^{-1})$. Additionally, in the each iteration, the following extra steps are needed for ω_k and η:

1. Sample ω_k from a $GIG(1/2,\, \beta_k^2,\, \eta^2)$.

2. Sample η^2 from a $Ga(a_{\eta^2} + d, b_{\eta^2} + 0.5 \sum_{k=1}^{d} \omega_k)$.

A.2 BQMR with fixed p_k

For simplicity of notation, denote $u_i = \sum_{k=1}^{K} k\xi_{ik}$ and let $A_i = A_{p_{u_i}}$, $B_i = B_{p_{u_i}}$, $C_i = C_{p_{u_i}}$, $\mu_i = \mu_{p_{u_i}}$, $\gamma_i = \gamma_{p_{u_i}}$. Then under a $N(\boldsymbol{\beta}_0,\, \Sigma_0)$ prior for $\boldsymbol{\beta}$, the posterior sampling algorithm consists of the following steps.

1. Sample $\boldsymbol{\beta}$ from $N(\boldsymbol{\beta}^*, \Sigma^*)$, where

$$\Sigma^* \;=\; \left(\sum_{i=1}^{n} \frac{\boldsymbol{x}_i \boldsymbol{x}_i^T}{B_i \sigma v_i} + \Sigma_0^{-1} \right)^{-1}, \quad \boldsymbol{\beta}^* \;=\; \Sigma^* \left\{ \sum_{i=1}^{n} \frac{\boldsymbol{x}_i [y_i - (\mu_i + \sigma C_i |\gamma_i| s_i + A_i v_i)]}{B_i \sigma v_i} + \Sigma_0^{-1} \boldsymbol{\beta}_0 \right\}$$

2. Sample v_i from a generalized inverse-Gaussian distribution, $GIG(1/2, a_i, b_i)$, where $a_i = [y_i - (\mu_i + \boldsymbol{x}_i^T \boldsymbol{\beta} + \sigma C_i |\gamma_i| s_i)]^2 / (B_i \sigma)$ and $b_i = 2/\sigma + A_i^2/(B_i \sigma)$, with $GIG(x \mid \nu, a, b) \propto x^{\nu - 1} \exp\{-0.5(a/x + bx)\}$, $a > 0$, $b > 0$.

3. Sample s_i from $N^+(\mu_{s_i}, \sigma_{s_i}^2)$, where $\sigma_{s_i}^2 = 1/[(C_i \gamma_i)^2 \sigma/(B_i v_i) + 1]$ and $\mu_{s_i} = \sigma_{s_i}^2 C_i |\gamma_i| [y_i - (\mu_i + \boldsymbol{x}_i^T \boldsymbol{\beta} + A_i v_i)]/(B_i v_i)$.

4. Sample σ from $GIG(a_\sigma + 1.5n, c, d)$, where $c = 2b_\sigma + 2\sum_{i=1}^n v_i + \sum_{i=1}^n [y_i - (\mu_i + \boldsymbol{x}_i^T \boldsymbol{\beta} + A_i v_i)]^2/(B_i v_i)$ and $d = \sum_{i=1}^n (C_i \gamma_i s_i)^2/(B_i v_i)$.

5. Sample μ_{p_k} sequentially from $N(\mu_{p_k}^*, (\sigma_{p_k}^*)^2) \mathbb{1}\{\mu_{p_{k-1}} < \mu_{p_k} < \mu_{p_{k+1}}\}$ for $k = 1, \ldots, K$, where $\mu_{p_0} = -\infty$, $\mu_{p_{K+1}} = +\infty$ and

$$(\sigma_{p_k}^*)^2 \;=\; \left[\frac{1}{\sigma_\mu^2} + \sum_{\xi_{ik}=1} \frac{1}{B_i \sigma v_i} \right]^{-1}, \quad \mu_{p_k}^* \;=\; (\sigma_{p_k}^*)^2 \sum_{\xi_{ik}=1} \frac{\boldsymbol{y}_i - (\boldsymbol{x}_i \boldsymbol{\beta} + \sigma C_i |\gamma_i| s_i + A_i v_i)}{B_i \sigma v_i}$$

6. Sample $\omega_1, \ldots, \omega_K$ from a $Dir(a_1^*, \ldots, a_K^*)$, where $a_k^* = a_k + \sum_{i=1}^n \mathbb{1}\{\xi_{ik} = 1\}$.

7. Sample $\xi_{i1}, \ldots, \xi_{iK}$ from a Multinomial distribution, where $p(\xi_{ik} = 1 | \boldsymbol{y}, \ldots) \propto \omega_k N(y_i \mid \mu_{p_k} + \boldsymbol{x}_i^T \boldsymbol{\beta} + \sigma C_{p_k} |\gamma_{p_k}| s_i + A_{p_k} v_i, \sigma B_{p_k} v_i)$.

8. Sample γ_{p_k} with a Metropolis-Hastings step. The full conditional of γ_{p_k} is,

$$p_{\gamma_{p_k}}(\gamma_{p_k} \mid \boldsymbol{y}, \ldots) \;\propto\; B_{p_k}^{-\frac{n_k}{2}} \exp\left\{ - \sum_{\xi_{ik}=1} \frac{[y_i - (\mu_{p_k} + \boldsymbol{x}_i^T \boldsymbol{\beta} + \sigma C_{p_k} |\gamma_{p_k}| s_i + A_{p_k} v_i)]^2}{2 B_{p_k} \sigma v_i} \right\}$$

where $n_k = \sum_{i=1}^n \mathbb{1}\{\xi_{ik} = 1\}$. We apply a logit transformation on γ_{p_k} to such that $\theta_{p_k} = \log[(\gamma_{p_k} - L_{p_k})/(U_{p_k} - \gamma_{p_k})]$. Then conditioning on all other parameters, there is,

$$p_{\theta_{p_k}}(\theta_{p_k} \mid \boldsymbol{y}, \ldots) \;\propto\; p_{\gamma_{p_k}}\left(\gamma_{p_k} = \frac{U_{p_k} e^{\theta_{p_k}} + L_{p_k}}{1 + e^{\theta_{p_k}}} \mid \boldsymbol{y}, \ldots \right) \cdot \frac{(U_{p_k} - L_{p_k}) e_{p_k}^{\theta}}{(1 + e_{p_k}^{\theta})^2}$$

We accept $\theta_{p_k}^*$ with probability $\min\{r_{p_k}, 1\}$, $r_{p_k} = p_{\theta_{p_k}}(\theta_{p_k}^* \mid \boldsymbol{y}, \ldots)/p_{\theta_{p_k}}(\theta_{p_k}^{(m-1)} \mid \boldsymbol{y}, \ldots)$.

If instead a Laplace prior is applied for $\boldsymbol{\beta}$ accompanied by a $Ga(a_{\eta^2}, b_{\eta^2})$ prior for η^2, then we replace the first step in the above algorithm by the following:

1. Sample $\boldsymbol{\beta}$ from $N(\boldsymbol{\beta}^*, \Sigma^*)$, where

$$\Sigma^* \quad = \quad \left(\sum_{i=1}^{n} \frac{\boldsymbol{x_i}\boldsymbol{x_i}^T}{B_i \sigma v_i} + \Omega \right)^{-1} \quad , \quad \boldsymbol{\beta}^* = \Sigma^* \left\{ \sum_{i=1}^{n} \frac{\boldsymbol{x}_i[y_i - (\mu_i + \sigma C_i |\gamma| \, s_i + A_i v_i)]}{B_i \sigma v_i} \right\}$$

where $\Omega = \text{diag}(\tau_1^{-1}, \ldots, \tau_d^{-1})$.

Additionally, in the each iteration, the following extra steps are needed for τ_j and η:

9. Sample τ_j from a $GIG(1/2,\, \beta_j^2,\, \eta^2)$.

10. Sample η^2 from a $Ga(a_{\eta^2} + d, b_{\eta^2} + 0.5 \sum_{j=1}^{d} \tau_j)$.

Adaptive Metropolis-Hastings

When we fit BQMR models with many quantile components, we apply the adaptive random walk Metropolis-within-Gibbs algorithm in Sherlock et al. (2010) to improve the efficiency of the Metropolis-Hastings steps for the shape parameters γ_{p_k}.

Following Sherlock et al. (2010), at the n-th iteration, we propose θ_{p_k} (γ_{p_k} after logit transformation) from the following jumping distribution,

$$\theta_{p_k} \quad \sim \quad \begin{cases} N(0, m_n^2 \tilde{\sigma}_n^2) & \text{with probability } 1 - \delta \\ N(0, \sigma_0^2) & \text{with probability } \delta \end{cases}$$

where δ is some small positive constant, e.g., 0.1, $\tilde{\sigma}_n^2$ is the variance of θ_{p_k} samples to date, and σ_0^2 is some fixed variance, which can be obtained from a single run of a non-adaptive Metropolis-Hastings. The scaling factor m_n is initialized to $m_0 = 2.38$, and the adaptation quantity is $\Delta = m_0/100$. If iteration n comes from the nonadaptive part of the proposal distribution, then $m_{n+1} = m_n$; otherwise:

- If the proposal was rejected, then $m_{n+1} = m_n - \Delta/\sqrt{n}$

- If the proposal was rejected, then $m_{n+1} = m_n + 2.3\Delta/\sqrt{n}$

This leads to an equilibrium acceptance rate of $1/(1+2.3) = 30\%$.

A.3 Semi-parametric BQMR with fixed p_k

The sampling scheme for the weights and μ_G is as follows:

1. Sample $\omega_1, \ldots, \omega_K$ from $Dir(a_1^*, \ldots, a_K^*)$, where $a_k^* = a_k + \sum_{i=1}^n \mathbb{1}\{\xi_{ik} = 1\}$ and

$$a_k = \begin{cases} \alpha_0 Be(\frac{p_k}{p_K} \mid \mu_G, \tau_G)\,, & k = 1 \\ \alpha_0[Be(\frac{p_k}{p_K} \mid \mu_G, \tau_G) - Be(\frac{p_{k-1}}{p_K} \mid \mu_G, \tau_G)]\,, & k = 2, \ldots, K-1 \\ \alpha_0[1 - Be(\frac{p_{k-1}}{p_K} \mid \mu_G, \tau_G)]\,, & k = K \end{cases}$$

2. Sample μ_G with a Metropolis-Hastings step. In the full conditional posterior,

$$p(\mu_G \mid \boldsymbol{y}, \ldots) \propto Dir(\omega_1, \ldots, \omega_K \mid a_1, \ldots, a_K) \cdot \mathbb{1}\left\{\mu_G \in (0,1)\right\}$$

where a_k takes the same form as in the previous step. We use a truncated normal distribution over $(0,1)$ as the jumping distribution with variance σ^2_{MH}, the tuning parameter. We accept the proposed sample ${\mu_G}^*$ with probability $\min\{r, 1\}$, where

$$r = \frac{p({\mu_G}^* \mid \boldsymbol{y}, \ldots)}{p(\mu_G \mid \boldsymbol{y}, \ldots)} \cdot \frac{TN(\mu_G \mid {\mu_G}^*, \sigma^2_{MH}, 0, 1)}{TN({\mu_G}^* \mid \mu_G, \sigma^2_{MH}, 0, 1)}$$

with $TN(\cdot)$ standards for the density of a truncated normal

A.4 BQMR with random p_k

For simplicity of notation, denote $u_i = \sum_{k=1}^K k\xi_{ik}$ and let $A_i = A_{p_{u_i}}$, $B_i = B_{p_{u_i}}$, $C_i = C_{p_{u_i}}$, $\mu_i = \mu_{p_{u_i}}$, $\gamma_i = \gamma_{p_{u_i}}$. Then under a Laplace prior for $\boldsymbol{\beta}$ accompanied by a $Ga(a_{\eta^2}, b_{\eta^2})$ prior for η^2, the posterior sampling algorithm consists of the following steps.

1. Sample $\boldsymbol{\beta}$ from $N(\boldsymbol{\beta}^*, \Sigma^*)$, where

$$\Sigma^* \quad = \quad \left(\sum_{i=1}^{n} \frac{\boldsymbol{x_i}\boldsymbol{x_i}^T}{B_i \sigma v_i} + \Omega\right)^{-1} \quad , \quad \boldsymbol{\beta}^* = \Sigma^* \left\{\sum_{i=1}^{n} \frac{\boldsymbol{x}_i[y_i - (\mu_i + A_i v_i)]}{B_i \sigma v_i}\right\}$$

   where $\Omega = \mathrm{diag}(\tau_1^{-1}, \ldots, \tau_d^{-1})$.

2. Sample v_i from a generalized inverse-Gaussian distribution, $GIG(1/2, a_i, b_i)$, where $a_i = [y_i - (\mu_i + \boldsymbol{x}_i^T\boldsymbol{\beta})]^2/(B_i\sigma)$ and $b_i = 2/\sigma + A_i^2/(B_i\sigma)$, with $GIG(x \mid \nu, a, b) \propto x^{\nu-1}\exp\{-0.5(a/x + bx)\}$, $a > 0$, $b > 0$.

3. Sample σ from $IG(a_\sigma + 1.5n, c)$, where $c = b_\sigma + \sum_{i=1}^n v_i + 0.5\sum_{i=1}^n [y_i - (\mu_i + \boldsymbol{x}_i^T\boldsymbol{\beta} + A_i v_i)]^2/(B_i v_i)$.

5. Sample μ_{p_k} sequentially from $N(\mu^*_{p_k}, (\sigma^*_{p_k})^2)\mathbb{1}\{\mu_{p_{k-1}} < \mu_{p_k} < \mu_{p_{k+1}}\}$ for $k = 1, \ldots, K$, where $\mu_{p_0} = -\infty$, $\mu_{p_{K+1}} = +\infty$ and

$$(\sigma^*_{p_k})^2 \quad = \quad \left[\frac{1}{\sigma_\mu^2} + \sum_{\xi_{ik}=1} \frac{1}{B_i\sigma v_i}\right]^{-1} \quad , \ \mu^*_{p_k} \quad = \quad (\sigma^*_{p_k})^2 \sum_{\xi_{ik}=1} \frac{\boldsymbol{y}_i - (\boldsymbol{x}_i\boldsymbol{\beta} + A_i v_i)}{B_i \sigma v_i}$$

6. Sample $\omega_1, \ldots, \omega_K$ from a $Dir(a_1^*, \ldots, a_K^*)$, where $a_k^* = a_k + \sum_{i=1}^n \mathbb{1}\{\xi_{ik} = 1\}$.

7. Sample $\xi_{i1}, \ldots, \xi_{iK}$ from a Multinomial distribution, where $p(\xi_{ik} = 1 | \boldsymbol{y}, \ldots) \propto \omega_k N(y_i \mid \mu_{p_k} + \boldsymbol{x}_i^T\boldsymbol{\beta} + A_{p_k} v_i, \sigma B_{p_k} v_i)$.

8. Sample p_k with a Metropolis-Hastings step. The full conditional of p_k is,

$$p(p_k \mid \boldsymbol{y}, \ldots) \quad \propto \quad \prod_{i=1}^{n} \Big[N(y_i \mid \mu_{p_k} + \boldsymbol{x}_i^T\boldsymbol{\beta} + A_{p_k} v_i, \sigma B_{p_k} v_i)\Big]^{\xi_{ik}} \mathbb{1}\{p_{k-1} < p_k < p_{k+1}\}$$

   We adopt a truncated normal distribution over $\left(p_{k-1}^{(m)}, p_{k+1}^{(m-1)}\right)$ with scale parameter σ_0 as the jumping distribution and accept the proposed p_k^* with probability $\min\{r_{p_k}, 1\}$, where

$$r_{p_k} \quad = \quad \frac{p(p_k^* \mid \boldsymbol{y}, \ldots)}{p(p_k^{(m-1)} \mid \boldsymbol{y}, \ldots)} \cdot \frac{TN(p_k^{(m-1)} \mid p_k^*, \sigma_0^2, p_{k-1}^{(m)}, p_{k+1}^{(m-1)})}{TN(p_k^* \mid p_k^{(m-1)}, \sigma_0^2, p_{k-1}^{(m)}, p_{k+1}^{(m-1)})}$$

   where $TN(\cdot)$ standards for the density of a truncated normal, $p_0 \equiv 0$ and $p_{K+1} \equiv 1$.

9. Sample τ_j from a $GIG(1/2,\ \beta_j^2,\ \eta^2)$.

10. Sample η^2 from a $Ga(a_{\eta^2} + d, b_{\eta^2} + 0.5 \sum_{j=1}^{d} \tau_j)$.

When we fit BQMR models with many quantile components, we apply the adaptive random walk Metropolis-within-Gibbs algorithm in Sherlock et al. (2010) the same way for the fixed-p_k algorithm in Appendix A.2.