New Bayesian Regression Models for Massive Data and Extreme Longitudinal Data

Yuanqi Chu

A thesis submitted for the degree of
Doctor of Philosophy

in the

Department of Mathematics
Brunel University London

May 2024

Abstract

The phenomena of heavy-tailedness and asymmetry are ubiquitous in a variety of practical applications. The intriguing property of heavy-tailedness implies that the underlying distribution is capable of producing anomalous observations which deviate too far from the main body of observations. Down-weighing such extreme observations for an asymmetric distribution can sacrifice inherited information and introduce considerable bias on parameter estimation. Over the past decades, two main approaches have emerged to tackle the distributional deviation caused by heavy-tailedness. The first procedure adopts mixture models to accommodate the heterogeneity in the distribution of the data. The second technique considers appropriate distributions to take care of the majority as well as the heavy tail of the data.

This thesis aims to make some novel contributions to the following three issues related to massive data and extreme longitudinal data exhibiting heavy-tailed characteristics. First, the multitude of existing literature coping with continuous distributions with heavy-tailedness contrasts sharply with the scarcity of integer-valued distributions. This especially applies to integer-valued time series modelling. Second, heavy tails can considerably shadow the nature of the dependence between the response and the covariates of interest, calling the normality assumption and conventional linear models into question. The quantile regression (QR) approach, which is robust to outlier contamination associated with heavy-tailed errors, serves as a remedy for these hurdles. Bayesian quantile regression (BQR) has received increasing attention from both theoretical and empirical viewpoints with wide applications and variants, but little attention has been paid to BQR for big data analysis. Third, the phenomena of heavy-tailedness often arise with the semi-continuous data, which are commonly characterized by a mixture of zero values and continuously distributed positive values. This conceptual framework leads to the formulation of a two-part model. The literature on two-part models, especially in Bayesian paradigms, for investigating quantiles of semi-continuous longitudinal data with bounded support such as the standard unit interval $(0, 1)$, is relatively limited.

This thesis encapsulates three themes to address the above-mentioned challenges: Bayesian integer-valued time series modelling with heavy-tailedness characteristics, Bayesian quantile regression for big data analysis and Bayesian

quantile parametric mixed regression for semi-continuous longitudinal data with bounded support. The main contributions are elaborated as below:

- Chapter 2 gives rise to the Bayesian inference for log-linear Beta–negative binomial integer-valued generalized autoregressive conditional heteroscedastic (BNB-INGARCH) models and conducts parameter estimations within adaptive Markov chain Monte Carlo frameworks. The conditions for the posterior distribution of the full model parameter to be proper given some general priors have been presented.

- Chapter 3 contributes to a new approach of Bayesian quantile regression for big data. This chapter introduces the structure link between Bayesian scale mixtures of normals linear regression and BQR via normal-inverse-gamma (NIG) distribution type of likelihood function, prior distribution and posterior distribution. The big data based algorithms for BQR and Bayesian LASSO quantile regression are provided and the proposed algorithms are demonstrated via simulations and a real-world data analysis.

- Chapter 4 introduces a two-part latent class Kumaraswamy quantile mixed regression with Bayesian inference for bounded longitudinal data that exhibit a large spike at zeros. Correlated random effects with class-specific covariance structures are formulated for the binary and the bounded positive components to account for both zero inflation and unobserved heterogeneity. The developed method portrays the trajectory of distinct latent class evolutions in the underlying outcome process, which provides valuable insights into the latent cluster structure at various quantiles encompassing the tails and caters to the exploration of skewed longitudinal data with bounded support.

Acknowledgements

First and foremost, I would like to express my deepest gratitude to my supervisor Prof. Keming Yu, for his invaluable mentorship and enduring support throughout my research endeavour. His expertise, patience, and constant encouragement have been instrumental in shaping the direction of this thesis and fostering my growth as a researcher. I am truly honoured and privileged to have had the opportunity to learn from him, and his dedication to excellence will resonate with me in my academic and professional pursuits.

I would like to express my heartfelt appreciation to the members of my thesis committee, Prof. Hongsheng Dai, my external examiner, Dr. Ben Parker, my internal examiner and Dr. David Bell, the independent chair, for their dedication to the viva process. Their invaluable insights, constructive comments and meticulous examination have been instrumental in shaping the academic rigor and quality of this research.

I am immensely grateful to the support and encouragement provided by the faculty and staff of the Department of Mathematics at Brunel University. Special gratitude goes to my co-supervisor and research development advisor, Dr. Diana Roman, whose expertise and assistance have been indispensable in navigating this research endeavor. I would like to extend my appreciation to Dr. Xian Wu from the Department of Mechanical and Aerospace Engineering, whose insightful discussions and encouragement have enriched my academic adventure.

It is with great esteem that I acknowledge the partial sponsorship I received from OptiRisk Systems company for my PhD research. Their financial support and giving me the permission to access the VIX futures tick count data analysed in Chapter 2 have been instrumental in the completion of this PhD thesis.

Finally, I am heavily indebted to my parents and my family. Their unwavering love, encouragement and support have been the pillars of strength throughout my academic journey.

List of Publications

i. Chu, Y., Yin, Z. and Yu, K. (2023). Bayesian scale mixtures of normals linear regression and Bayesian quantile regression with big data and variable selection. *Journal of Computational and Applied Mathematics* 428, p. 115192.

ii. Chu, Y. and Yu, K. (2023). Bayesian log-linear beta-negative binomial integer-valued Garch model. *Computational Statistics*, pp. 1-20.

iii. Chu, Y., Hu, X. and Yu, K. (2021, September). Bayesian Quantile Regression for Big Data Analysis. In *Interational Conference on Bayesian Young Statistician Meeting* (pp. 11-22). Cham: Springer International Publishing.

Contents

Chapter 1

Introduction

1.1 Heavy-tailedness phenomena and asymmetry

The distributional assumption on disturbances is the most important premise in probabilistically modelling in a diverse set of disciplines including physical, biological, and social science. Theoretical supports such as estimator consistency and statistical hypotheses formulation depend crucially on the distributional assumption for the underlying model. The traditional and most widely employed distributional assumption in statistical procedures has been the Gaussian distribution. The Gaussian theory works well especially when fitting the empirical data which conform to a bell-shaped curve and the random fluctuations in the experimental observations of some quantity are scattered symmetrically around the true value of the quantity. A pivotal statistical argument for the Gaussian assumption is the Central Limit Theorem (CLT), which establishes that the arithmetic mean of a number of independent and identically distributed (*i.i.d.*) random variables, drawn with overall mean and finite variance, approaches the Gaussian distribution as the number of variables diverges to infinity. The fundamental theory of the Gaussian probability law manifests its prevalence in all branches of science discipline dealing with disturbances and randomness.

However, empirical studies which provide evidence against the Gaussian assumption have proliferated in literature. The phenomena of heavy-tailedness and asymmetry are ubiquitous in a variety of practical applications. It is well known that financial returns often exhibit fat tails which makes heavy-tail analysis aimed at improving risk management indispensable. Other fields where a prevalence of heavy tails has emerged encapsulate meteorology, physics, hydrology and engineering (see, Resnick, 1997; Lu and Molz, 2001; Cappé et al., 2002; Kysely and Picek, 2007; Frankenberg et al., 2016, among others). The intriguing property of heavy-tailedness implies that the underlying distribution is capable of producing anomalous data points which deviate too far from the main body of observations. Modelling data with an asymmetric heavy-tailed distribution merits well-deserved attention through its connection with the concept of outliers. Outliers occur in the tails of a sample, prompted by deviations from the bulk of the data conspicuously exceeding expectancy. The existence of outliers can dramatically impact the results of statistical analysis and interpretation. When the distribution is symmetric around the mean, the problem caused by outliers can be remedied by robust statistical techniques (Huber, 2011; Rousseeuw and Hubert, 2011; Maronna et al., 2019), which aim at removing or putting less weights on anomalous data objects to preserve the integrity of the models. However, down-weighing outliers for an asymmetric distribution, where departures from the mean arise from a long tail of larger or smaller values in the underlying process, can sacrifice some inherited information and introduce considerable bias on parameter estimation. This sketches the significant relevance of heavy-tailed phenomena and underlines the importance of having the right probabilistic distribution models for capturing their behaviours. Over the past decades, two main approaches have emerged to tackle the distributional deviation caused by heavy-tailednes. The first procedure incorporates mixture models to accommodate the heterogeneity in the distribution of the data. The parameter estimates for such models are commonly implemented via expectation–maximization (EM) approach or Markov chain Monte Carlo (MCMC) sampling algorithm. The second technique is to adopt appropriate distributions to take care of the majority as well as the heavy tail of the data.

The search for continuous distributions with the nature of heavy-tailedness has been invigorated in literature. Early explorations in this area can be traced back to Lévy (1925), where the Lévy stable distributions with support $(-\infty, \infty)$ replying on a power law behaviour were introduced to represent a rich class of probability distributions allowing for skewness and fat tails. Other proposed distributions that share the same attractive power law tails and real line support incorporate the Pearson-type IV (e.g., Nagahara, 1999), Student t (e.g., Zabell, 2008), doubly Pareto-uniform (DPU) distribution (Singh et al., 2007) and among others. The study of modelling heavy-tailed continuous data restricted to a bounded interval has also grasped an increasing attention. Aban et al. (2006) investigated the truncated Pareto distribution with bounded support and illustrated the derived maximum likelihood estimators (MLEs) facilitate robust tail estimation for truncated models with power law tails. Hahn (2008) proposed the beta-rectangular (BR) distribution as a mixture of the uniform and the classical beta distribution. The BR distribution permits modelling heavy-tailed bounded data in equal proportions of both tails. To circumvent this limitation, very recently, Figueroa-Zúñiga et al. (2022) introduced the trapezoidal beta (TB) distribution which extends both the beta and rectangular beta distributions to permit the description of bounded data with heavy right or left tails in different proportions.

The multitude of existing literature coping with continuous distributions with heavy-tailedness stands in sharp contrast to the relative scarcity of research on integer-valued distributions. This especially applies to integer-valued time series modelling. Time series analysis can be considered as a subclass of longitudinal studies, with a specific focus on the univariate data observed over multiple time points. The challenge in modelling heavy-tailed discrete time series data lies in seeking appropriate distributions to accommodating data with tails longer than the customarily employed negative binomial and Poisson distributions. The well-known Poisson inverse Gaussian (PIG) distribution has long served as a robust alternative to the negative binomial (Holla, 1967; Willmot, 1987). Barreto-Souza (2019) proposed an integer-valued autoregressive (INAR) process with PIG distribution for overdispersed count time series data. Nevertheless, heavy-tailedness was not referred to by the author. Silva and Barreto-Souza (2019) explored a general class of integer-valued generalized autoregressive conditional heteroscedastic (INGARCH) models based on a flexible family of mixed Poisson (MP) distributions, which incorporate the PIG and Poisson generalized hyperbolic secant processes. The authors illustrated the robustness of the PIG-INGARCH model than the negative binomial count process considered. Recently, Qian et al. (2020) proposed a new INAR(1) process with the generalized Poisson inverse Gaussian (GPIG) innovations. The developed model can take into account both equidispersion and overdispersion, as well as excessive zero characteristics in count time series data. Gorgi (2020) explored a heavy-tailed mixture of negative binomial distributions, known as the Beta–negative binomial (BNB) distribution, and introduced a BNB autoregression process for modelling integer-valued time series with outliers, where a linear observation-driven dynamic equation for the conditional mean has been specified. The above-mentioned contributions all performed model parameter estimations from the frequentist perspective.

Conventional regression approaches fit models replying on certain central tendencies, including the mean, median and mode. Ordinary least square (OLS) regression follows the central tendency theorem and possesses readily comprehensible statistical properties for conditional expectations. This fact contributes to the primacy of OLS regression as an empirical tool. Nevertheless, towards extreme distributions the OLS model loses its effectiveness to investigate the nuanced relationships between response and explanatory variables. Quantile regression (QR) has emerged as a robust and distribution-free modelling tool since the seminal work of Koenker and Bassett (1978a). Part of the appeal of quantile regression evolves from a natural parallel with the conventional ordinary least square regression. The OLS estimates provide the minimum mean-squared error linear approximation to a conditional expectation function (White, 1980). On the other hand, QR is anchored on minimizing the sum of absolute residuals and provides a way to model the conditional quantiles of a response variable with respect to the explanatory variables. The quantile regression model is an extremely powerful tool for capturing a more complete picture of the entire conditional distribution than traditional linear regression. In fact, the pervasive phenomena in natural and social science, such as skewness, heavy tails, heteroskedasticity and truncated and censored data, can considerably shadow the nature of the dependence between the response and the covariates of interest, calling the normality assumption and conventional linear models into question. The quantile regression approach serves as a remedy for these hurdles. The nature of QR implies that it is robust to outlier contamination associated with heavy-tailed errors, which has become a subject of intense investigation and attracted considerable research interest in the past decades (see, e.g., Zou and Yuan, 2008; Fan et al., 2014; Sherwood and Wang, 2016; Huang et al., 2017; Chen et al., 2020, among others).

In the "big data" era of statistical science, the richness and complexity of data structures along with the increase of extreme values and heterogeneity may see quantile regression methods more competent than mean regressions to dig deep into the data and grab information from it. In particular, with the advanced power of computer technology, complicated quantile regression-based models could be developed under a Bayesian framework, and Bayesian quantile regression (BQR) has received increasing attention from both theoretical and empirical viewpoints with wide applications and variants (see Kozumi and Kobayashi, 2011a; Bernardi et al., 2015; Wang et al., 2016b; Rodrigues and Fan, 2017; Petrella and Raponi, 2019; Gonçalves et al., 2020 and among others). So far, in the context of quantile regression, several methods have been explored for big data analysis (Wu and Yin, 2015; Yu et al., 2017; Gu et al., 2018; Chen et al., 2019b and among others), but little attention has been paid to such methodology under Bayesian inference paradigm.

Heavy tails can also be ascribed to data with an abundance of zeros, where the shifting of the overall sample mean closer to zero will increase the skewness of the observations (Lord and Geedipally, 2018). Integer-valued outcomes are typically modelled employing discrete distributions such as the Poisson or negative binomial distribution. However, one frequent manifestation of heterogeneity in count data is that the incidence of zeros is greater than what a standard count distribution would predict. In this way, flexible mixture mechanisms are usually adopted to accommodate the unique characteristic of the data. Zero-inflated models (Lambert, 1992) are mixture models of two data generation processes, where one generates the point mass at zero and the other allows for both zero and non-zero counts. In contrast, hurdle models (Mullahy, 1986) can be viewed as a two-component mixture model consisting of a zero mass and a positive component for observations, where the latter follows a truncated count distribution. A similar phenomenon arises with semi-continuous data, which are commonly characterized by a mixture of zero values and continuously distributed positive values. The semi-continuous variables can be regarded as arising from two distinct stochastic processes, where one determines the occurrence of zero values and the other identifies the actual values for nonzero observations. This conceptual framework leads to the formulation of a two-part model, which caters to both the preponderance of zeros and the typically highly skewed distribution of nonzero observations (Heckman, 1976). A lognormal distribution is frequently considered to model the nonzero continuous component, giving rise to the Bernoulli-lognormal two-part model (Cragg, 1971). Alternative distributions have been investigated to relax the log-symmetry condition inherent in the lognormal distribution, such as the two-part log-skew-normal model (Chai and Bailey, 2008) and the two-part generalized gamma model (Liu et al., 2010).

The existing literature on two-part models targeted at semi-continuous data where the continuous component is characterized by positive support is well explored. On the other hand, the literature encompassing two-part models for variables with bounded support is relatively limited. Practitioners oftentimes encounter variables whose values fall into the standard unit interval $(0, 1)$, such as rates, proportions and concentration indices. The most commonly used model with such variables is the beta regression model proposed by Ferrari and Cribari-Neto (2004). The chief motivation for the beta regression model lies in the flexibility delivered by the postulated beta law. However, Kumaraswamy (1976) noted that the beta law may fail to deliver desirable fits especially when the data are hydrological observations of small frequency. He then introduced a new distribution referred to as the Kumaraswamy distribution. The Kumaraswamy distribution is very similar to the beta distribution but has the advantage of possessing a closed form for both the probability density function and the cumulative distribution function. For details on the Kumaraswamy law, one can refer to Jones (2009). Bayes et al. (2017) proposed a new quantile parametric mixed regression model which is built upon a reparameterization of the Kumaraswamy distribution in terms of a given quantile and the precision parameter, and formulated a Bayesian approach for parameter inference including model comparison criteria. The authors demonstrated that the developed quantile parametric model complements for bounded response variables modelling such as the poverty index in Brazilian municipalities.

The Kumarawasmy and beta distributions cannot be employed when the data contain zeros and/or ones. New laws embodying a discrete component which imposes positive probability to such point(s) have been proposed. Ospina and Ferrari (2010) considered inflated beta distributions for modelling fractional data. More recently, Cribari-Neto and Santos (2019) introduced inflated Kumaraswamy distributions. The authors employed the standard Kumaraswamy parametrization indexed by two shape parameters. On the contrary, Bayer et al. (2021) developed an inflated version of the Kumaraswamy distribution utilizing the median-based parametrization from Mitnik and Baek (2013) and conducted model parameter estimation by maximum likelihood inference.

1.2 Bayesian analysis

Bayesian analysis offers unparalleled flexibility in accommodating hierarchical structures inherent in many real-world datasets. Unlike classical frequentist methods, which often struggle with the complexity of hierarchical models, Bayesian frameworks seamlessly integrate prior information and incorporate varying levels of uncertainty across parameters (Gelman et al., 2013). This flexibility is paramount in capturing intricate relationships within data hierarchies, especially in scenarios where traditional models fall short.

By leveraging prior distributions, researchers can encode existing information into the analysis, thereby enhancing the robustness of inference. This aspect becomes particularly valuable in situations where data are scarce or noisy, allowing for more informed and stable estimations compared to classical methods. A significant challenge encountered in certain statistical models, such as those involving dispersion parameter estimation in distributions like the negative binomial, is the presence of likelihood multimodality. Traditional maximum likelihood estimation may falter in such scenarios, often converging to local optima (Dai et al., 2013). As discussed in Section 2.4, the dispersion parameter r within our developed log-linear BNB-INGARCH model framework faces a challenge in achieving unique identification. Consequently, the likelihood function may present with multiple maxima or regions of elevated likelihood, introducing ambiguity in parameter estimation. This ambiguity increases the risk of erratic estimates characterized by small fluctuations or the emergence of multiple modes. This inherent complexity underscores the importance of careful consideration and robust methodologies in addressing parameter estimation challenges within the model. Bayesian methods offer distinct advantages in addressing these complexities by regularizing the estimation process through the incorporation of prior knowledge about the parameter in Bayesian inference. This proves particularly beneficial when the available data alone may be insufficient to resolve multiple modes, thereby facilitating more robust and comprehensive parameter estimation and overcoming the limitations of classical approaches.

When tackling mixture models through the frequentist approach, challenges such as non-convergence or entrapment in local modes can impede effective inference. Consequently, Bayesian analysis emerges as a compelling alternative, offering robustness against these pitfalls. One notable advantage lies in its accommodation of uncertainty regarding the number of classes present in the data, a common scenario in practical applications. In such instances, where the number of classes is unknown or the model lacks identifiability due to limited sample size, Bayesian methodology shines by enabling the incorporation of prior information. Even with scant data, Bayesian inference provides meaningful estimates through the formulation of posterior distributions for model parameters, underscoring its suitability for research endeavors characterized by complex mixture models and limited data availability.

Bayesian methods also excel in the realm of latent class analysis, providing a principled framework for modelling intricate data structures and capturing latent heterogeneity (Frühwirth-Schnatter, 2006). By harnessing Bayesian inference, researchers can seamlessly integrate correlated random effects with class-specific covariance structures, thereby delineating distinct latent class evolutions and gaining profound insights into underlying cluster structures across different quantiles. This comprehensive approach facilitates the exploration of skewed longitudinal data, enabling a nuanced understanding of the underlying dynamics that extends beyond the capabilities of classical frequentist methods (Gelman et al., 2013). Thus, Bayesian approaches offer a powerful toolkit for addressing the complexities inherent in modern data analysis tasks, enabling researchers to extract meaningful insights from diverse and heterogeneous datasets (Gelman et al., 2013).

1.3 Thesis structure

This thesis consists of three main chapters.

Chapter 2 gives rise to the Bayesian inference for log-linear Beta–negative binomial integer-valued generalized autoregressive conditional heteroscedastic (BNB-INGARCH) models and conducts parameter estimations within adaptive Markov chain Monte Carlo frameworks. Moreover, the conditions for the posterior distribution of the full model parameter to be proper given some general priors have been derived and presented. The empirical application on high-frequency intraday VIX futures tick count data indicates that the proposed model is adequate and provides better performance than its counterpart INGARCH models under negative binomial distribution assumptions. This chapter complements another aspect of current literature on modelling discrete time series with heavy-tailedness characteristics.

Chapter 3 contributes to a new approach of Bayesian quantile regression for big data. This chapter introduces the structure link between Bayesian scale mixtures of normals linear regression and Bayesian quantile regression via normal-inverse-gamma (NIG) distribution type of likelihood function, prior distribution and posterior distribution. The big data based algorithms for BQR and Bayesian LASSO quantile regression are provided and the proposed algorithms are demonstrated via simulations and two real-world data analyses.

Chapter 4 develops a two-part latent class Kumaraswamy quantile mixed regression with Bayesian inference for bounded longitudinal data that exhibit a large spike at zeros. The binomial component is specified via mixed effects probit regression and the continuous component is formulated through a Kumaraswamy quantile mixed effects model. Correlated random effects with class-specific covariance structures are established for the binary and the bounded positive mechanisms to account for both zero inflation and unobserved heterogeneity. The presented method portrays the trajectory of distinct latent class evolutions in the underlying outcome process, which provides valuable insights into the latent cluster structure at various quantiles encompassing the tails and caters to the exploration of skewed longitudinal data with bounded support.

The methodologies presented in this thesis are unified by their shared focus on addressing challenges inherent in analysing massive data sets and extreme longitudinal data. While diverse in their specific applications, these chapters are linked by a common thread of leveraging Bayesian inference to develop robust statistical models capable of capturing complex data structures and providing insightful interpretations.

By adopting Bayesian inference, we embrace a flexible framework that accommodates uncertainty in parameter estimation, facilitates incorporation of prior knowledge and enables principled handling of complex data structures. Unlike frequentist methods, which often rely on asymptotic approximations and assume fixed parameters, Bayesian approaches offer a coherent framework for quantifying uncertainty and updating beliefs as new data become available.

The integration of Bayesian techniques across the chapters underscores their versatility and adaptability to diverse modelling contexts. From log-linear Beta–negative binomial INGARCH models for count data to Bayesian quantile regression for big data sets and latent class Kumaraswamy quantile mixed regression for bounded longitudinal data, the Bayesian paradigm offers a cohesive toolkit for addressing the multifaceted challenges posed by modern data analysis.

As a whole, the methods proposed in this thesis offer several advantages. Firstly, they provide robust solutions for modelling data with heavy-tailed distributions, skewed distributions and bounded support, which are common characteristics of many real-world data sets. Secondly, integrating prior information facilitates enhanced parameter estimation, particularly in scenarios characterized by restricted data availability. Thirdly, these models boast adaptability, accommodating diverse data structures and research inquiries, thereby rendering them applicable across a spectrum of domains encompassing finance, transportation, energy and healthcare.

The overarching contribution of this thesis lies in advancing the state-of-the-art in statistical modelling for complex data scenarios. By introducing novel Bayesian regression models tailored to address specific challenges such as heavy-tailedness, zero inflation and extreme longitudinal dynamics, this work expands the methodological toolkit available to researchers and practitioners. Furthermore, the empirical validations and real-world applications demonstrate the practical utility and effectiveness of the proposed approaches, underscoring their potential to yield valuable insights and inform decision-making in diverse domains. Through these contributions, this thesis serves as a foundational step towards enhancing our ability to extract meaningful information from massive and complex data sets, thereby facilitating more accurate and interpretable scientific inference across diverse domains.

Chapter 2

Bayesian log-linear Beta–negative binomial integer-valued GARCH model

2.1 Introduction

Despite a long history in the literature analysing continuous time series variables, it is only in the recent years or so that much attention has been given to time series variables that are integer-valued (see Davis et al., 2016; Fahrmeir et al., 2013; Weiß, 2018; Winkelmann, 2008, and references therein for reviews). Generally, integer-valued time series models can be classified into two categories: 'thinning' operator based models (Scotto et al., 2015) and regression based models (Fokianos, 2012; Tjøstheim, 2016).

To allow for dependence between time series data, two classes of models have been proposed in Cox et al. (1981): observation-driven models and parameter-driven models. In observation-driven models, the mean of the conditional distribution of the current observation y_t is directly specified as a function of past observations $y_{t-1}, \ldots, y_1$. In parameter-driven models, dependence among observations is introduced via latent factors which follow a stochastic process, such as a hidden Markov chain (Leroux and Puterman, 1992), or a latent stationary auto-regressive process (Chan and Ledolter, 1995; Zeger, 1988). Compared with parameter-driven models, observation-driven models are easier to fit in practical contexts with numerous covariates and long time series. See Zeger and Qaqish (1988) for a review of various observation-driven models for count time series data. A reference for the substantial development of observation-driven models can be found in Kedem and Fokianos (2005). A variety of observation-driven models for count responses have been developed. Davis et al. (2003, 2005) presented a class of observation-driven models for time series of Poisson counts and provided properties of the maximum likelihood estimators. Ahmad and Francq (2016) derived regularity conditions for the consistency and asymptotic normality (CAN) of the Poisson quasi-maximum likelihood estimator (QMLE) for time series of counts. However, the equidispersion assumption in Poisson distribution makes it too restrictive to be applied in empirical settings. Given this, Drescher (2005) considered various generalized count distributions for observation driven models and explored their maximum likelihood estimations. Regarding existing R packages (R Core Team, 2021), the **glarma** package (Dunsmuir and Scott, 2015) provides functions for estimation, testing, diagnostic checking and forecasting based on the generalized linear autoregressive moving average (GLARMA) class of observation-driven models for discrete-valued time series with regression variables.

The benchmark parameter-driven count data model introduced in Zeger (1988) has been widely extended. It has been considered as a class of the state-space model, which extends the generalized linear model by introducing a latent autoregressive process as the conditional mean function. The parameter-driven models allow the distribution of y_t to be dependent on this latent process and can deal with auto-correlation as well as over-dispersion in the model. However, parameter estimations in parameter-driven models require considerable computational effort. The main issue lies in the calculation requirement of very high dimensional integrals when using maximum likelihood estimation techniques, such that estimation methods based on Monte-Carlo (MC) integration are typically considered. To estimate the parameters of parameter-driven models, Chan and Ledolter (1995) employed a Monte Carlo EM algorithm. Kuk and Cheng (1997) considered the MC Newton Raphson method. However, such estimation approaches are not yet routinely available and therefore not ready for general use (Davis et al., 2003).

Bayesian estimation for time series of counts turns out to be a feasible and more elaborate alternative. Applications of Bayesian paradigm to count times series have mainly focused on parameter-driven models. Dynamic latent factor models within Bayesian count time series contexts have been actively studied (see, e.g., Chib and Winkelmann, 2001; Durbin and Koopman, 2000; West and Harrison, 2006). Hay and Pettitt (2001) presented a fully Bayesian analysis of counts time series for a parameter-driven model with the form of a generalized linear mixed model, and investigated its application to the control of an infectious disease. Unlike the MC EM estimation approach for parameter-driven models, the Markov chain Monte Carlo (MCMC) procedure provides information of posterior distributions for both regression and time series parameters. In maximum likelihood-based estimations, estimation uncertainty is produced by constructing confidence intervals around the point forecasts. However, this kind of confidence intervals can only be justified asymptotically. When counts are small, such approximation is less accurate and the Bayesian technique arises as a prime candidate. When forecasting counts from the Bayesian perspective, not only the parameter uncertainty, but also the uncertainty caused by model specification, can be directly incorporated into the predictive probability mass function, which is a natural outcome of Bayes' theorem.

Although parameter-driven models are very flexible, the existence of unobservable latent factors brings a heavy computational burden even manageable via the MCMC sampling technique. On the other hand, Bayeisan analysis of the more parsimonious observation-driven models has received growing attention recently. Generalized autoregressive moving average (GARMA) model extends the univariate Gaussian ARMA model to a flexible non-Gaussian observation-driven model. Silveira de Andrade B et al. (2015) investigated the Bayesian approach for GARMA models with Poisson, binomial and negative binomial distributions. They utilized the Bayesian model selection criteria to choose the most appropriate model. Another advantage of Bayesian methodology over the corresponding frequentist procedure for forecasting discrete time series data is that Bayesian approach can produce only integer estimates of the count variable, while traditional forecasting often yields non-coherent (i.e. non-integer) estimates. For example, the Autoregressive Integrated Moving Average (ARIMA) model is one of the most prominent methods in financial time series forecasting. It has shown robust and efficient capability for short-term predictions and has been extensively applied to economics and finance fields (Contreras et al., 2003; Khashei et al., 2009; Lee and Ko, 2011, among others). However, forecasts from ARIMA model can give negative values. Techniques such as log scale transformation or constrained forecast might guarantee non-negative predictions, but with the burden of elaborate post-processing and consequences of obtaining back-transformed forecasts that behave abnormally. Given the fact that many actual data in socioeconomic and business areas cannot have negative values, the classical ARIMA forecasting methods are improper when applied to non-negative count data series. From another point of view, Bayesians utilize the likelihood and prior multiplicity to generate forecasts from posterior predictive distributions by the iterative loop of MCMC procedures. Therefore, Bayesian model is prone to obey the non-negative value rules with its probabilistic predictive distributions, providing new perceptions for time series forecasting research. Nariswari and Pudjihastuti (2019) implemented Bayesian time series estimation on ARIMA model for monthly medicine demand count data, showing the validity of Bayesian time series approach to avoid negative-value predictions, which is consistent with characteristics in the actual medicine data where the stock cannot have a negative value. McCabe and Martin (2005) developed Bayesian predictions of low count time series within the context of the integer-valued first-order autoregressive (INAR(1)) class of model, and showed the Bayesian method is feasible for producing coherent forecasts. Estimation uncertainty associated with both parameters and model specification is fully incorporated in their proposed methodology.

A commonly used model in most count time series data is the Poisson integer-valued generalized autoregressive conditional heteroscedastic (Poisson INGARCH) model proposed in Ferland et al. (2006). Since then, this model has been widely explored (see Doukhan et al., 2012; Neumann, 2011, among others). However, the Poisson INGARCH model is not eligible to be applied in existence of potential extreme observations due to its equidispersion assumption. To this end, Zhu (2011) developed a negative binomial (NB) INGARCH model via the maximum likelihood approach. The NB-INGARCH model is flexible and allows for both overdispersion and extreme observations simultaneously. Later, Christou and Fokianos (2014) explored probabilistic properties and quasi-likelihood estimation for NB-INGARCH(1,1) process, and Xiong and Zhu (2019) considered a robust quasi-likelihood estimation for this process with an application to transaction counts. From a Bayesian perspective, Truong et al. (2017) proposed a hysteretic Poisson INGARCH model within the MCMC sampling scheme to estimate model parameters and adopted the Bayesian information criteria for model comparison. They highlighted their proposed model with a better performance of hysteresis in modelling the integer-valued time series. Chen et al. (2019a) developed a Markov switching Poisson INGARCH model within a Bayesian framework to cope with the lagged dependence, overdispersion, consecutive zeros, non-linear dynamics and time-varying coefficients for the meteorological variables. Some studies considered the natural candidates for the Poisson model. Chen and Lee (2017) proposed a Bayesian causality test based on the Poisson, negative binomial and log-linear Poisson INGARCH models with applications to climate and crime data. Recently, Chen and Khamthong (2020) introduced two nonlinear negative binomial INGARCH models (Markov switching and threshold specifications) along with the exogenous covariates in the conditional mean to describe time series of counts. They conducted parameter estimations and one-step-ahead forecasting via the Bayesian MCMC methods.

When modelling time series with outlying and atypical data, a commonly used approach is to develop models based on heavy-tailed distributions. The literature coping with continuous-valued time series with extreme observations is well explored via the Student's t-distribution (Harvey and Luati, 2014). However, current literature on modelling discrete time series with heavy-tailedness is less considered. To fill this gap, very recently, Qian et al. (2020) proposed a new integer-valued autoregressive process with generalized Poisson-inverse Gaussian (GPIG) innovations to model heavy-tailed count time series data. Gorgi (2020) explored a heavy-tailed mixture of negative binomial distributions, known as the Beta–negative binomial (BNB) distribution, and developed a BNB autoregression process for modelling integer-valued time series with outliers, where a linear observation-driven dynamic equation for the conditional mean has been specified. Both Qian et al. (2020) and Gorgi (2020) employed maximum likelihood approaches to estimate the model parameters. This chapter gives rise to the Bayesian inference for log-linear BNB-INGARCH models and conducts parameter estimations within adaptive Markov chain Monte Carlo frameworks. Moreover, the conditions for the posterior distribution of the full model parameter to be proper given some general priors have been derived and presented.

2.2 The log-linear BNB-INGARCH model

The Beta-negative binomial distribution can be represented as a beta mixture of the negative binomial distribution. Denote a discrete random variable Y, then $Y \sim BNB(\beta, r, \alpha)$ if its probability mass function (PMF) is given by

$$f(Y = y) = \frac{\Gamma(y+r)}{\Gamma(y+1)\Gamma(r)} \frac{B(\alpha + r, \beta + y)}{B(\alpha, \beta)}, y \in \mathcal{N} \tag{2.1}$$

where $\Gamma(\cdot)$ is the gamma function and $B(\cdot, \cdot)$ is the beta function. $r > 0$ is the dispersion parameter, $\alpha > 0$ is the tail parameter and $\beta > 0$. We follow the parameterization of the BNB distribution in terms of its mean parameter λ presented in Gorgi (2020)

$$f(Y = y) = \frac{\Gamma(y+r)}{\Gamma(y+1)\Gamma(r)} \frac{B(\alpha + r, \frac{(\alpha-1)\lambda}{r} + y)}{B(\alpha, \frac{(\alpha-1)\lambda}{r})}, y \in \mathcal{N} \tag{2.2}$$

with mean $\lambda > 0$, dispersion $r > 0$ and tail $\alpha > 1$. Let $\{y_t\}, t = 1, \ldots, n$ denote a univariate time series with the conditional distribution following the representation of $BNB(\lambda_t, r, \alpha)$ at time t. We model the log-intensity process $\mu_{t+1} = \log(\lambda_{t+1})$ in terms of a linear auto-regression process lying on its own past μ_t and the past observation y_t. The log-linear BNB integer-valued generalized auto-regressive conditional heteroscedastic model of order (1,1) is defined by

$$\begin{aligned} y_t|\mathcal{F}_{t-1} &\sim BNB(\lambda_t, r, \alpha), \\ \mu_{t+1} &= \log(\lambda_{t+1}), \\ \mu_{t+1} &= \omega + \phi \log(y_t + 1) + \tau \mu_t \end{aligned} \tag{2.3}$$

where $\mathcal{F}_{t-1}$ denotes the "σ-field" generated by $\{y_{t-1}, y_{t-2}, \ldots\}$. Here we follow Fokianos and Tjøstheim (2011) to choose $\log(y_t + c)$ with constant $c = 1$ in model (2.3) to map zeros of y_t into zeros of $\log(y_t + 1)$. Other reasonable choices for c may be considered. Note that model (2.3) accommodates both negative and positive serial correlations by allowing parameters ω, ϕ, τ to take values in $\mathbb{R}$, whereas the linear BNB-INGARCH model introduced by Gorgi (2020) accommodates positive serial correlation only by restricting parameters to be positive to guarantee positivity of the conditional mean. Moreover, model (2.3) permits faster increase or decrease in λ_t according to the values of ω, ϕ and τ than the linear model. Extensions to higher order log-linear BNB-INGARCH (p, q) models can be given as follows:

$$\begin{aligned} y_t|\mathcal{F}_{t-1} &\sim BNB(\lambda_t, r, \alpha), \\ \mu_{t+1} &= \log(\lambda_{t+1}), \\ \mu_{t+1} &= \omega + \sum_{i=1}^{p} \phi_i \log(y_{t+1-i} + 1) + \sum_{j=1}^{q} \tau_j \mu_{t+1-j} \end{aligned} \tag{2.4}$$

We note that the BNB distribution belongs to the class of mixed Poisson distributions and can approximate arbitrarily well the negative binomial distribution as well as the Poisson distribution (Gorgi, 2020; Johnson et al., 2005; Wang, 2011). Specifically, as $\alpha \to \infty$, the parameterized distribution $\text{BNB}(\lambda, r, \alpha)$ converges to a NB distribution with dispersion r and success probability $\lambda/(\lambda + r)$. Furthermore, as $r \to \infty$, the BNB converges to a Poisson distribution with mean λ (Gorgi, 2020). Given this Poisson approximation to the BNB distribution, we follow Douc et al. (2013, Lemma 14) and Liboschik et al. (2017) to impose the conditions $\{|\phi|, |\tau| < 1, |\phi + \tau| < 1\}$ and $\{|\phi_i|, |\tau_j| < 1, |\sum_{i=1}^{p} \phi_i + \sum_{j=1}^{q} \tau_j| < 1\}$ to guarantee stationarity of the proposed processes (2.3) and (2.4) respectively. We further follow Wang et al. (2014) and Gorgi (2020) to preassign the initial point λ_1 to a fixed positive value for both models (2.3) and (2.4). As noted in Gorgi (2020), this approach is quite standard in the literature of observation-driven time series models. In fact, Gorgi (2020) showed that the filtered parameter $\{\hat{\lambda}_t(\boldsymbol{\theta})\}_{t\in\mathbb{N}}$ converges exponentially almost surely and uniformly over the compact parameter sets Θ to a unique stationary and ergodic sequence $\{\tilde{\lambda}_t(\boldsymbol{\theta})\}_{t\in\mathbb{Z}}$ for any initialization $\hat{\lambda}_1(\boldsymbol{\theta}) \in \mathbb{R}^+$.

2.3 Bayesian inference

Due to the high computational demand, we resort to Bayesian analysis for parameter estimations and inferences of the log-linear BNB-INGARCH processes. Without loss of generality, we focus on the first-order specification ($p = q = 1$) as presented in model (2.3) for simplicity of inference illustration. We denote the time series of interest as $\mathbf{y} = (y_1, y_2, \ldots, y_n)^T$ and $\boldsymbol{\theta} = (\omega, \phi, \tau, r, \alpha)^T$ as the entire unknown parameter vector. By the Bayes theorem, the posterior distribution $p(\boldsymbol{\theta}|\mathbf{y})$ is given by

$$p(\boldsymbol{\theta}|\mathbf{y}) \propto L(\mathbf{y}|\boldsymbol{\theta})\pi(\boldsymbol{\theta}) \tag{2.5}$$

where $\pi(\boldsymbol{\theta})$ denotes the prior distribution and $L(\mathbf{y}|\boldsymbol{\theta})$ represents the likelihood function

$$L(\boldsymbol{y}|\boldsymbol{\theta}) = \prod_{t=2}^{n} \frac{\Gamma(y_t + r)}{\Gamma(r)} \frac{B\left(\alpha + r, \frac{(\alpha-1)e^{\omega+\phi\log(y_{t-1}+1)+\tau\log(\lambda_{t-1})}}{r} + y_t\right)}{B\left(\alpha, \frac{(\alpha-1)e^{\omega+\phi\log(y_{t-1}+1)+\tau\log(\lambda_{t-1})}}{r}\right)} \tag{2.6}$$

with $\omega \in \mathbb{R}$, $|\phi| < 1$, $|\tau| < 1$, $|\phi + \tau| < 1$, $r > 0$ and $\alpha > 1$. We implement Bayesian inference for parameter groups (i) ω; (ii) $\{\phi, \tau\}$; (iii) r; and (iv) α with the assumption that they are priori-independent. The conditional posterior distributions for each parameter group are presented as follows. Hereafter let θ_j denote the j-th element of the parameter vector $\boldsymbol{\theta}$, with $j = 1, 2, 3, 4$ referring to ω, $\{\phi, \tau\}$, r and α respectively, and $\boldsymbol{\theta}_{-j}$ denote the vector of all parameters excluding the component θ_j.

(i) For the intercept ω, we consider a normal prior $\pi(\omega) = N(\mu_\omega, \sigma_\omega^2)$. The full conditional posterior distribution of ω is then as follows:

$$\begin{aligned} p(\omega|\boldsymbol{\theta}_{-1}, \mathbf{y}) &\propto L(\mathbf{y}|\boldsymbol{\theta})\pi(\omega) \\ &\propto L(\mathbf{y}|\boldsymbol{\theta}) e^{-\frac{(\omega-\mu_\omega)^2}{2\sigma_\omega^2}} \end{aligned} \tag{2.7}$$

(ii) For the prior $\pi(\phi, \tau)$ of the parameter block $\{\phi, \tau\}$, we employ constrained normals $\pi(\phi) = N(\mu_\phi, \sigma_\phi^2)$ and $\pi(\tau) = N(\mu_\tau, \sigma_\tau^2)$ with $\{\phi, \tau\}$ satisfying the set A

$$|\phi| < 1, |\tau| < 1, |\phi + \tau| < 1 \tag{2.8}$$

Then the full conditional posterior distribution of $\{\phi, \tau\}$ is given by

$$\begin{aligned} p(\phi, \tau|\boldsymbol{\theta}_{-2}, \mathbf{y}) &\propto L(\mathbf{y}|\boldsymbol{\theta})\pi(\phi, \tau) \\ &\propto L(\mathbf{y}|\boldsymbol{\theta}) e^{-\frac{(\phi-\mu_\phi)^2}{2\sigma_\phi^2}} e^{-\frac{(\tau-\mu_\tau)^2}{2\sigma_\tau^2}} \mathcal{I}(A) \end{aligned} \tag{2.9}$$

where $\mathcal{I}(\cdot)$ denotes the indicator function.

(iii) For the dispersion parameter r, we impose a gamma prior $\pi(r) = \mathrm{Ga}(\eta_{1r}, \eta_{2r})$ with shape parameter η_{1r} and rate parameter η_{2r}, then the full conditional posterior distribution of r is given by

$$\begin{aligned} p(r|\boldsymbol{\theta}_{-3}, \mathbf{y}) &\propto L(\mathbf{y}|\boldsymbol{\theta})\pi(r) \\ &\propto L(\mathbf{y}|\boldsymbol{\theta}) r^{\eta_{1r}-1} e^{-\eta_{2r}\cdot r} \end{aligned} \tag{2.10}$$

(iv) For the tail parameter α, we impose a truncated gamma prior $\pi(\alpha) = \mathrm{Ga}(\eta_{1\alpha}, \eta_{2\alpha})\mathcal{I}(\alpha > 1)$ with shape $\eta_{1\alpha}$ and rate $\eta_{2\alpha}$, then the full conditional posterior distribution of α is given by

$$\begin{aligned} p(\alpha|\boldsymbol{\theta}_{-4}, \mathbf{y}) &\propto L(\mathbf{y}|\boldsymbol{\theta})\pi(\alpha) \\ &\propto L(\mathbf{y}|\boldsymbol{\theta}) \alpha^{\eta_{1\alpha}-1} e^{-\eta_{2\alpha}\cdot\alpha} \mathcal{I}(\alpha > 1) \end{aligned} \tag{2.11}$$

Theorem 2.1 elaborated below presents the sufficient conditions for the posterior distribution of $\boldsymbol{\theta}$ to be proper given some general priors.

Theorem 2.1. *Let $\{y_t|\mathcal{F}_{t-1}\}_{t\in\mathbb{Z}^+}$ denote the target count time series with the conditional Beta-negative binomial distribution and $\boldsymbol{\theta} = (\omega, \phi, \tau, r, \alpha)^T$ be the full parameter vector. For ease of notation, we denote $\kappa(\boldsymbol{\theta}) = (\alpha - 1)/r \cdot e^{\omega+\phi\log(y_{t-1}+1)+\tau\log(\lambda_{t-1})}$ in the remainder of this section. Then under model (2.3) and proper prior specifications, the posterior distribution of $\boldsymbol{\theta}$ given $\mathbf{y} = (y_1, y_2, \ldots, y_n)^T$ is obtained by*

$$p(\boldsymbol{\theta}|\boldsymbol{y}) \propto \prod_{t=2}^{n} \frac{\Gamma(y_t + r)}{\Gamma(r)} \frac{B(\alpha + r, \kappa(\boldsymbol{\theta}) + y_t)}{B(\alpha, \kappa(\boldsymbol{\theta}))} \pi(\omega)\pi(\phi, \tau)\pi(\alpha)\pi(r)$$

and is well defined.

Proof. Under proper prior specifications, we have

$$
\begin{aligned}
&\int p(\boldsymbol{\theta}|\boldsymbol{y})d\boldsymbol{\theta} \\
&\propto \int \left\{ \prod_{t=2}^{n} \frac{\Gamma\left(y_t + r\right)}{\Gamma(r)} \frac{B\left(\alpha + r, \kappa(\boldsymbol{\theta}) + y_t\right)}{B\left(\alpha, \kappa(\boldsymbol{\theta})\right)} \times \pi(\omega)\pi(\phi,\tau)\pi(\alpha)\pi(r) \right\} d\omega\, d\phi\, d\tau\, d\alpha\, dr \\
&= \int \left\{ \prod_{t=2}^{n} \frac{\Gamma\left(y_t + r\right)}{\Gamma(r)} \frac{\Gamma(\alpha + r)}{\Gamma(\alpha)} \frac{\Gamma\left(\kappa(\boldsymbol{\theta}) + y_t\right)}{\Gamma\left(\kappa(\boldsymbol{\theta})\right)} \frac{\Gamma\left(\alpha + \kappa(\boldsymbol{\theta})\right)}{\Gamma\left(\alpha + \kappa(\boldsymbol{\theta}) + y_t + r\right)} \pi(\omega)\pi(\phi,\tau) \right. \\
&\left. \pi(\alpha)\pi(r) \right\} d\omega\, d\phi\, d\tau\, d\alpha\, dr \qquad (2.12)
\end{aligned}
$$

where the equation follows by the relation $B(a,b) = \frac{\Gamma(a)\Gamma(b)}{\Gamma(a+b)}$. From the Stirling's approximation of the gamma function $\Gamma(\cdot)$, the following approximation is obtained for large r and α:

$$
\begin{aligned}
(2.12) &\approx \int r^{\sum_{t=2}^{n} y_t} \left\{ \prod_{t=2}^{n} \frac{\Gamma(\alpha + r)}{\Gamma(\alpha)} \frac{\Gamma\left(\kappa(\boldsymbol{\theta}) + y_t\right)}{\Gamma\left(\kappa(\boldsymbol{\theta})\right)} \frac{\Gamma\left(\alpha + \kappa(\boldsymbol{\theta})\right)}{\Gamma\left(\alpha + \kappa(\boldsymbol{\theta}) + y_t + r\right)} \right. \\
&\quad \left. \pi(\omega)\pi(\phi,\tau)\pi(\alpha)\pi(r) \right\} d\omega\, d\phi\, d\tau\, d\alpha\, dr \\
&\approx \int r^{\sum_{t=2}^{n} y_t} \left\{ \prod_{t=2}^{n} \alpha^{r} \Big(\kappa(\boldsymbol{\theta})\Big)^{y_t} \Big(\alpha + \kappa(\boldsymbol{\theta})\Big)^{-(y_t + r)} \pi(\omega)\pi(\phi,\tau)\pi(\alpha)\pi(r) \right\} \\
&\quad d\omega\, d\phi\, d\tau\, d\alpha\, dr \\
&= \int r^{\sum_{t=2}^{n} y_t} \left\{ \prod_{t=2}^{n} \left(\frac{\alpha}{\alpha + \kappa(\boldsymbol{\theta})} \right)^{r} \left(\frac{\kappa(\boldsymbol{\theta})}{\alpha + \kappa(\boldsymbol{\theta})} \right)^{y_t} \pi(\omega)\pi(\phi,\tau)\pi(\alpha)\pi(r) \right\} \\
&\quad d\omega\, d\phi\, d\tau\, d\alpha\, dr \\
&= \int r^{\sum_{t=2}^{n} y_t} \left\{ \prod_{t=2}^{n} \left(\frac{1}{1 + \frac{\kappa(\boldsymbol{\theta})}{\alpha}} \right)^{r} \left(\frac{1}{1 + \frac{\alpha}{\kappa(\boldsymbol{\theta})}} \right)^{y_t} \pi(\omega)\pi(\phi,\tau)\pi(\alpha)\pi(r) \right\} \\
&\quad d\omega\, d\phi\, d\tau\, d\alpha\, dr \qquad (2.13)
\end{aligned}
$$

For $r \geqslant 1$, we obtain

$$
\begin{aligned}
(2.13) &\leqslant \int r^{\sum_{t=2}^{n} y_t} \left\{ \prod_{t=2}^{n} \frac{1}{1 + \frac{r\kappa(\boldsymbol{\theta})}{\alpha}} \left(\frac{1}{1 + \frac{\alpha}{\kappa(\boldsymbol{\theta})}} \right)^{y_t} \pi(\omega)\pi(\phi,\tau)\pi(\alpha)\pi(r) \right\} \\
&\quad d\omega\, d\phi\, d\tau\, d\alpha\, dr \\
&\leqslant \int r^{\sum_{t=2}^{n} y_t} \left\{ \prod_{t=2}^{n} \frac{1}{1 + \frac{r\kappa(\boldsymbol{\theta})}{\alpha}} \frac{1}{1 + \frac{\alpha y_t}{\kappa(\boldsymbol{\theta})}} \pi(\omega)\pi(\phi,\tau)\pi(\alpha)\pi(r) \right\} \\
&\quad d\omega\, d\phi\, d\tau\, d\alpha\, dr \qquad (2.14)
\end{aligned}
$$

where the first inequality follows by Bernoulli's inequality $(1+x)^d \geqslant 1+dx$ for real numbers $d \geqslant 1, x \geqslant -1$ and the second inequality follows by Bernoulli's inequality $(1+x)^d \geqslant 1+dx$ for integer $d \geqslant 0$ and real number $x \geqslant -1$. Without loss of generality, we assume that the first n_1 y_t's have zero-valued observations and the remaining $(n-n_1)$ observations have positive y_t's. Then we have

$$
\begin{aligned}
(2.14) &= \int r^{\sum_{t=n_1+1}^{n} y_t} \Bigg\{ \prod_{t=2}^{n_1} \frac{1}{1+\left(\frac{\alpha-1}{\alpha}\right) e^{\omega+\tau \log(\lambda_{t-1})}} \prod_{t=n_1+1}^{n} \frac{1}{1+\frac{r\kappa(\boldsymbol{\theta})}{\alpha}} \frac{1}{1+\frac{\alpha y_t}{\kappa(\boldsymbol{\theta})}} \\
&\quad \pi(\omega)\pi(\phi,\tau)\pi(\alpha)\pi(r) \Bigg\} \, d\omega \, d\phi \, d\tau \, d\alpha \, dr \\
&\leqslant \int r^{\sum_{t=n_1+1}^{n} y_t} \Bigg\{ \prod_{t=2}^{n_1} \frac{1}{\left(\frac{\alpha-1}{\alpha}\right) e^{\omega+\tau \log(\lambda_{t-1})}} \prod_{t=n_1+1}^{n} \frac{\alpha}{r\kappa(\boldsymbol{\theta})} \frac{\kappa(\boldsymbol{\theta})}{\alpha y_t} \pi(\omega)\pi(\phi,\tau) \\
&\quad \pi(\alpha)\pi(r) \Bigg\} \, d\omega \, d\phi \, d\tau \, d\alpha \, dr \\
&= \int r^{\sum_{t=n_1+1}^{n} y_t} \Bigg\{ \prod_{t=2}^{n_1} \frac{\alpha \, e^{-\{\omega+\tau \log(\lambda_{t-1})\}}}{\alpha-1} \prod_{t=n_1+1}^{n} \frac{\alpha \, e^{-\{\omega+\phi \log(y_{t-1}+1)+\tau \log(\lambda_{t-1})\}}}{\alpha-1} \\
&\quad \frac{(\alpha-1) \, e^{\omega+\phi \log(y_{t-1}+1)+\tau \log(\lambda_{t-1})}}{\alpha \, r \, y_t} \pi(\omega)\pi(\phi,\tau)\pi(\alpha)\pi(r) \Bigg\} \, d\omega \, d\phi \, d\tau \, d\alpha \, dr \\
&= C_1 \int r^{\sum_{t=n_1+1}^{n} (y_t-1)} \left(\frac{\alpha}{\alpha-1}\right)^{n_1-1} e^{-(n_1-1)\omega} e^{-(\sum_{t=2}^{n_1} \log(\lambda_{t-1}))\tau} \\
&\quad \pi(\omega)\pi(\phi,\tau)\pi(\alpha)\pi(r) \, d\omega \, d\phi \, d\tau \, d\alpha \, dr
\end{aligned} \tag{2.15}
$$

where $C_1 = \prod_{t=n_1+1}^{n} \frac{1}{y_t}$. By the Binomial approximation $\left(\frac{\alpha}{\alpha-1}\right)^{n_1-1} = \left(1+\frac{1}{\alpha-1}\right)^{n_1-1} \approx 1+\frac{n_1-1}{\alpha-1}$ for large α and further assuming $\alpha \geqslant M$ for any large positive number M, we obtain

$$
\begin{aligned}
(2.15) &\approx C_1 \int r^{\sum_{t=n_1+1}^{n} (y_t-1)} \left(1+\frac{n_1-1}{\alpha-1}\right) e^{-(n_1-1)\omega} e^{-(\sum_{t=2}^{n_1} \log(\lambda_{t-1}))\tau} \\
&\quad \pi(\omega)\pi(\phi,\tau)\pi(\alpha)\pi(r) \, d\omega \, d\phi \, d\tau \, d\alpha \, dr \\
&\leqslant C_1 \int r^{\sum_{t=n_1+1}^{n} (y_t-1)} \left(1+\frac{n_1-1}{M-1}\right) e^{-(n_1-1)\omega} e^{-(\sum_{t=2}^{n_1} \log(\lambda_{t-1}))\tau} \\
&\quad \pi(\omega)\pi(\phi,\tau)\pi(\alpha)\pi(r) \, d\omega \, d\phi \, d\tau \, d\alpha \, dr \\
&= C_2 \int r^{\sum_{t=n_1+1}^{n} (y_t-1)} e^{-(n_1-1)\omega} \pi(\omega)\pi(r) \, d\omega \, dr < \infty
\end{aligned}
$$

given that the prior density functions $\pi(\omega), \pi(\phi,\tau), \pi(\alpha)$ and $\pi(r)$ integrate to a finite quantity, the $\sum_{t=n_1+1}^{n} (y_t-1)$-th moment about the origin of r exists and the moment generating function $M_\omega(1-n_1)$ of ω exists. Here C_2 is a constant unrelated with the parameters of interest. □

It follows from the proof of Theorem 2.1 that the priors for (ϕ, τ) and α can be chosen very flexibly. In fact, any proper distributions can be considered because the appropriateness of the posterior will not be affected. On the other hand, the choice of the prior $\pi(r)$ requires existence of the $\sum_{t=n_1+1}^{n}(y_t - 1)$-th moment about the origin of r and the choice of the prior $\pi(\omega)$ requires existence of the moment generating function $M_\omega(1-n_1)$.

Since the obtained posterior distributions do not correspond to closed-form distributions, we resort to MCMC sampling methods. Specifically, for faster convergence and better mixing of the chain, we follow the adaptive MCMC method of Chen and So (2006). We employ the random-walk Metropolis-Hastings in the first H iterations (the burn-in period) and the independent-kernel Metropolis-Hastings in the following $(N-H)$ iterations to draw samples for the parameter vector $\boldsymbol{\theta}$. The adaptive MCMC procedure is provided as follows:

Step1. Set initial values for $\boldsymbol{\theta}^{(0)} = (\omega^{(0)}, \phi^{(0)}, \tau^{(0)}, r^{(0)}, \alpha^{(0)})^T$.

Step2. When $1 \leqslant k \leqslant H$, we adopt the random-walk Metropolis-Hastings algorithm for sampling $\boldsymbol{\theta}^{(k)}$:

Step2.1. Generate candidate values $\boldsymbol{\theta}^* = \{\theta_l^*\}$, $l = 1, 2, \ldots, 5$, where $\theta_l^* = \theta_l^{(k-1)} + \epsilon, \epsilon \sim N(0, \sigma_l^2)$, and the tuning parameter σ_l^2 is selected to achieve the acceptance rate around 23%.

Step2.2. Keep the candidate values if $\boldsymbol{\theta}^*$ satisfies that: $\theta_1 > 0$, $\{\theta_2, \theta_3\}$ do not violate the stationarity conditions of the model, $\theta_4 > 0$ and $\theta_5 > 1$. Otherwise, go back to Step 2.1.

Step2.3. Calculate the acceptance probability

$$\text{Prob}(\boldsymbol{\theta}^*, \boldsymbol{\theta}^{(k-1)}) = \min\left\{1, \frac{p(\boldsymbol{\theta}^*)}{p(\boldsymbol{\theta}^{(k-1)})}\right\},$$

where $p(\cdot)$ is the target posterior distribution given in (2.5). Then generate a random uniform number $u \in [0,1]$. If $u < \text{Prob}(\boldsymbol{\theta}^*, \boldsymbol{\theta}^{(k-1)})$, accept the new candidate and set $\boldsymbol{\theta}^{(k)} = \boldsymbol{\theta}^*$. Otherwise, set $\boldsymbol{\theta}^{(k)} = \boldsymbol{\theta}^{(k-1)}$.

Step3. When $k \geqslant H+1$, we adopt the independent-kernel Metropolis-Hastings algorithm for sampling $\boldsymbol{\theta}^{(k)}$:

Step3.1. Generate candidate values $\boldsymbol{\theta}^* \sim N(\boldsymbol{\mu}_\theta, \boldsymbol{\Omega}_\theta)$, where the sample mean $\boldsymbol{\mu}_\theta$ and sample covariance matrix $\boldsymbol{\Omega}_\theta$ are calculated using the burn-in H iteration samples.

Step3.2. Keep the candidate values if $\boldsymbol{\theta}^*$ satisfies that: $\theta_1 > 0$, $\{\theta_2, \theta_3\}$ do not violate the stationarity conditions of the model, $\theta_4 > 0$ and $\theta_5 > 1$. Otherwise, go back to Step 3.1.

Step3.3. Calculate the acceptance probability

$$\mathrm{Prob}(\boldsymbol{\theta}^*, \boldsymbol{\theta}^{(k-1)}) = \min\left\{1, \frac{p(\boldsymbol{\theta}^*)g(\boldsymbol{\theta}^{(k-1)})}{p(\boldsymbol{\theta}^{(k-1)})g(\boldsymbol{\theta}^*)}\right\},$$

where $g(\cdot)$ is the Gaussian proposal density with mean $\boldsymbol{\mu}_\theta$ and sample covariance matrix $\boldsymbol{\Omega}_\theta$. Then generate a random uniform number $u \in [0, 1]$. If $u < \mathrm{Prob}(\boldsymbol{\theta}^*, \boldsymbol{\theta}^{(k-1)})$, accept the new candidate and set $\boldsymbol{\theta}^{(k)} = \boldsymbol{\theta}^*$. Otherwise, set $\boldsymbol{\theta}^{(k)} = \boldsymbol{\theta}^{(k-1)}$.

Step4. Go to the next iteration or stop if the chain has converged.

Given the obtained N iteration samples $\boldsymbol{\theta}^{(1)}, \boldsymbol{\theta}^{(2)}, \ldots, \boldsymbol{\theta}^{(N)}$, we discard the first H in the burn-in period and perform model parameter estimations using the remaining $(N - H)$ iterations.

2.4 Simulation analysis

To examine the performance of the adaptive MCMC algorithm for the log-linear BNB-INGARCH model, we conduct a simulation analysis under the following data generating process (DGP) with sample sizes $n = 100$ and $n = 250$. The count series y_t is sampled from the log-linear BNB-INGARCH model (2.3) where $(\omega, \phi, \tau, r, \alpha)^T$ is set to be $(0.65, 0.7, -0.2, 5, 3)^T$. To investigate the sensitivity of the prior and hyperparameter selections, we consider the following five prior calibrations for the parameters of interest:

Prior 1: $\omega \sim N(0.1, 0.3^2)$, $\{\phi, \tau\} \sim N(0.1, 0.25^2) \cdot N(0.1, 0.75^2)\mathcal{I}(A)$, $r \sim$ Ga $(10, 0.5)$ and $\alpha \sim \mathrm{Ga}(10, 1)\mathcal{I}(\alpha > 1)$
Prior 2: $\omega \sim N(0.1, 0.5^2)$, $\pi(\phi, \tau) \propto \mathcal{I}(A)$, $r \sim \mathrm{Ga}(10, 0.5)$ and $\alpha \sim \mathrm{Exp}(0.1)$ $\mathcal{I}(\alpha > 1)$
Prior 3: $\omega \sim N(0.1, 0.5^2)$, $\pi(\phi, \tau) \propto \mathcal{I}(A)$, $r \sim \mathrm{Ga}(10, 0.5)$ and $\alpha \sim \mathcal{IG}(0.5, 1)$ $\mathcal{I}(\alpha > 1)$
Prior 4: $\omega \sim N(0.1, 0.5^2)$, $\pi(\phi, \tau) \propto \mathcal{I}(A)$, $r \sim \mathrm{Exp}(0.01)(r > 0)$ and $\alpha \sim \mathcal{LN}$ $(0, 1.25)\mathcal{I}(\alpha > 1)$
Prior 5: $\omega \sim N(0.1, 0.3^2)$, $\{\phi, \tau\} \sim N(0.1, 0.25^2) \cdot N(0.1, 0.75^2)\mathcal{I}(A)$, $r \sim$ Ga $(10, 1)$ and $\pi(\alpha) \propto \frac{1}{(1+\alpha)^2}\mathcal{I}(\alpha > 1)$

where $\mathrm{Exp}(\eta)$ denotes the exponential distribution with rate parameter η, $\mathcal{IG}(\rho_1, \rho_2)$ represents the inverse gamma distribution with shape ρ_1 and scale ρ_2, $\mathcal{LN}(\delta, \zeta)$ indicates the log-normal distribution with parameters mean δ and standard deviation ζ of the distribution on the natural log scale, and the set A is given in (2.8). Note that in Priors 2-4, we employ a constrained uniform prior on $\{\phi, \tau\}$ which configures a flat prior on the parameters restricted by the indicator $\mathcal{I}(A)$. We perform 10,000 MCMC iterations by discarding 5000 iterations as a burn-in sample for inference in each scenario. We employ R (R Core Team, 2021) as our programming language for all computational tasks. The

CPU time required for parameter estimation in each prior and sample size configuration is under 1.3 hrs over 100 replications. Table 2.1 presents the averages of posterior means (Ave Mean), medians (Ave Median), standard deviations (Ave Std), and 95% credible intervals (95% CI) for all model parameters. For a comparison with the frequentist approach, we report the simulation results for the maximum likelihood (ML) estimator of the considered model in Table 2.2. The averages of the mean and the standard deviation for different parameter values and sample sizes are obtained from 100 Monte Carlo replications.

To facilitate the ML estimation, we follow Gorgi (2020) to reparameterize r and α in terms of their inverse. Our empirical findings reveal that, particularly in small sample sizes, a situation may arise where the likelihood function exhibits a flat profile in the vicinity of the true parameter values. As a consequence, the estimates for r and α could become arbitrarily large and unbounded. To mitigate this issue, we employ inverse reparameterizations based upon the parameterization invariance property of maximum likelihood estimates (same ML estimation obtained independent of the chosen parametrization) (Barndorff-Nielsen, 1983) to secure stable parameter estimates for r and α, and to enhance the convergence properties of optimization algorithms.

Given Tables 2.1 and 2.2, we summarize the simulation results as follows:

(1) We observe that the employed adaptive MCMC approach gives a reasonably accurate estimate of the parameters of interest for small sample sizes. The standard deviation decreases as the sample size increases from 100 to 250 in both Bayesian and ML estimation scenarios. Furthermore, all the reported standard deviation values for parameters ω, ϕ and τ under Bayesian estimation are smaller than that of ML estimator in each simulation scheme.

(2) The simulation results demonstrate that both positive and negative serial correlations can be captured by the proposed model. Moreover, the average posterior means and posterior medians are overall reasonably close to the true values of the parameters, implying the validity of the considered adaptive MCMC method. We therefore suggest using posterior medians as the model parameter estimates since the median is a robust measure of central tendency compared with the mean.

(3) We observe that the adaptive MCMC procedure is robust to the selection of priors and hyperparameters via delivering similar reasonably accurate estimation results under different setting scenarios.

(4) We extensively examine the sensitivity of starting values by specifying different starting points for the adaptive MCMC sampler. We also investigate the sensitivity relating to the choice of λ_1 by randomly setting different initialization values for the intensity process. We observe that the employed MCMC sampler is robust to the selection of the starting values and the initial intensity

as different calibrations deliver similar reasonably accurate estimation results for the considered small sample sizes n.

Figure 2.1 displays the kernel density of the ML estimates. The density estimates affirm the consistency of the estimations, as the distributions converge towards the true parameter values with an augmentation in sample size. The graphical representation underscores that particularly in scenarios with limited sample sizes, one can observe more skewed distributions and subtle fluctuations in the estimates of the tail parameter α, along with the emergence of bimodal distribution patterns for the dispersion r. This aligns with the Lemma 8 presented in Gorgi (2020), which states that the dispersion parameter may not be uniquely identified and its distribution is anticipated to exhibit a second lower mode in small sample sizes. The obtained results support the notion that, particularly in scenarios where ML estimation encounters challenges such as skewed distributions, small sample sizes and multiple maxima in the likelihood function, Bayesian inference emerges as a more robust and coherent framework for parameter estimation. In situations with limited sample sizes, there is an inherent elevation of uncertainty surrounding parameter estimates. Furthermore, when parameters lack unique identification, the likelihood function may manifest multiple maxima or regions of high likelihood, resulting in ambiguity during parameter estimation and the potential for erratic estimates with small bumps or multiple modes. Bayesian methods offer distinct advantages in addressing these complexities. The incorporation of prior knowledge about the parameter in Bayesian inference serves to regularize the estimation process, proving particularly beneficial when the available data alone may be inadequate to resolve multiple modes. Bayesian approaches, notably through MCMC methods, possess the capability to comprehensively explore the entire parameter space. This feature mitigates the risk of becoming trapped in a local maximum, a concern inherent in maximum likelihood inference. The ability of Bayesian estimation to provide a full distribution of parameter estimates allows for the quantification of uncertainty and enhances the reliability and interpretability of parameter estimates.

Parameter	True	Ave	Ave	Ave	95% CI			Ave	Ave	Ave	95% CI	
	Value	Mean	Median	Std	P2.5	P97.5		Mean	Median	Std	P2.5	P97.5
$n = 100$	Prior 1						$n = 250$	Prior 1				
ω	0.65	0.5658	0.5669	0.1617	0.2814	0.8633	ω	0.5627	0.5551	0.1317	0.3459	0.8200
ϕ	0.7	0.7693	0.7735	0.0970	0.5833	0.9293	ϕ	0.7504	0.7521	0.0703	0.6222	0.8744
τ	-0.2	-0.2063	-0.2143	0.1392	-0.4410	0.0585	τ	-0.2422	-0.2419	0.0968	-0.4150	-0.0691
r	5	5.1305	5.2343	0.5883	3.8319	5.8679	r	4.5433	4.6227	0.4698	3.4939	5.1462
α	3	3.3755	3.4244	0.2565	2.8188	3.7010	α	3.1450	3.1641	0.1995	2.7072	3.4282
$n = 100$	Prior 2						$n = 250$	Prior 2				
ω	0.65	0.6949	0.6835	0.2183	0.3175	1.1202	ω	0.6240	0.6186	0.1647	0.3375	0.9459
ϕ	0.7	0.8659	0.8751	0.0794	0.6967	0.9800	ϕ	0.8067	0.8101	0.0771	0.6604	0.9417
τ	-0.2	-0.3547	-0.3551	0.1208	-0.5722	-0.1178	τ	-0.3223	-0.3237	0.0986	-0.5010	-0.1365
r	5	5.8206	5.9774	0.7867	4.0849	6.8480	r	4.9661	5.0983	0.6215	3.5263	5.7409
α	3	3.4463	3.4812	0.3754	2.6580	3.9898	α	3.1395	3.1633	0.2855	2.5597	3.5751
$n = 100$	Prior 3						$n = 250$	Prior 3				
ω	0.65	0.6650	0.6696	0.1975	0.3245	1.0129	ω	0.6178	0.6201	0.1501	0.3488	0.8956
ϕ	0.7	0.8284	0.8428	0.0757	0.6677	0.9318	ϕ	0.7982	0.8035	0.0725	0.6603	0.9157
τ	-0.2	-0.2330	-0.2469	0.0989	-0.3800	-0.0396	τ	-0.2653	-0.2744	0.0776	-0.3924	-0.1106
r	5	4.6396	4.7453	0.5302	3.5318	5.3350	r	4.5329	4.5921	0.4615	3.5893	5.2177
α	3	2.8958	2.9108	0.2756	2.3527	3.3170	α	2.8689	2.8884	0.2136	2.4275	3.1926
$n = 100$	Prior 4						$n = 250$	Prior 4				
ω	0.65	0.6820	0.6768	0.2048	0.3123	1.0290	ω	0.6046	0.6036	0.1390	0.3471	0.8563
ϕ	0.7	0.8110	0.8208	0.0813	0.6434	0.9292	ϕ	0.7727	0.7750	0.0709	0.6357	0.8895
τ	-0.2	-0.2324	-0.2438	0.1021	-0.3921	-0.0227	τ	-0.2270	-0.2346	0.0738	-0.3380	-0.0712
r	5	5.4730	5.5326	0.6845	4.0962	6.4899	r	5.1184	5.2356	0.5467	3.9067	5.8666
α	3	3.1024	3.1083	0.4257	2.3294	3.8078	α	2.9992	3.0056	0.2918	2.4664	3.4866
$n = 100$	Prior 5						$n = 250$	Prior 5				
ω	0.65	0.6133	0.6081	0.1841	0.2898	0.9537	ω	0.6294	0.6235	0.1441	0.3743	0.9170
ϕ	0.7	0.7372	0.7444	0.1146	0.5151	0.9372	ϕ	0.7320	0.7304	0.0785	0.5933	0.8788
τ	-0.2	-0.1452	-0.1622	0.1555	-0.4058	0.1589	τ	-0.2122	-0.2117	0.1074	-0.4116	-0.0189
r	5	4.8441	4.9470	0.7813	3.2057	5.9008	r	4.2510	4.3176	0.6038	2.9771	5.1067
α	3	2.8488	2.8851	0.2583	2.3006	3.2031	α	2.7703	2.7972	0.2056	2.3320	3.0563

Prior 1: $\omega \sim N(0.1, 0.3^2)$, $\{\phi, \tau\} \sim N(0.1, 0.25^2) \cdot N(0.1, 0.75^2)\mathcal{I}(A)$, $r \sim \text{Ga}(10, 0.5)$ and $\alpha \sim \text{Ga}(10, 1)\mathcal{I}(\alpha > 1)$
Prior 2: $\omega \sim N(0.1, 0.5^2)$, $\pi(\phi, \tau) \propto \mathcal{I}(A)$, $r \sim \text{Ga}(10, 0.5)$ and $\alpha \sim \text{Exp}(0.1)\mathcal{I}(\alpha > 1)$
Prior 3: $\omega \sim N(0.1, 0.5^2)$, $\pi(\phi, \tau) \propto \mathcal{I}(A)$, $r \sim \text{Ga}(10, 0.5)$ and $\alpha \sim \mathcal{IG}(0.5, 1)\mathcal{I}(\alpha > 1)$
Prior 4: $\omega \sim N(0.1, 0.5^2)$, $\pi(\phi, \tau) \propto \mathcal{I}(A)$, $r \sim \text{Exp}(0.01)(r > 0)$ and $\alpha \sim \mathcal{LN}(0, 1.25)\mathcal{I}(\alpha > 1)$
Prior 5: $\omega \sim N(0.1, 0.3^2)$, $\{\phi, \tau\} \sim N(0.1, 0.25^2) \cdot N(0.1, 0.75^2)\mathcal{I}(A)$, $r \sim \text{Ga}(10, 1)$ and $\pi(\alpha) \propto \frac{1}{(1+\alpha)^2}\mathcal{I}(\alpha > 1)$

Table 2.1 Simulation results for the Bayesian log-linear BNB-INGARCH model obtained from 100 replications

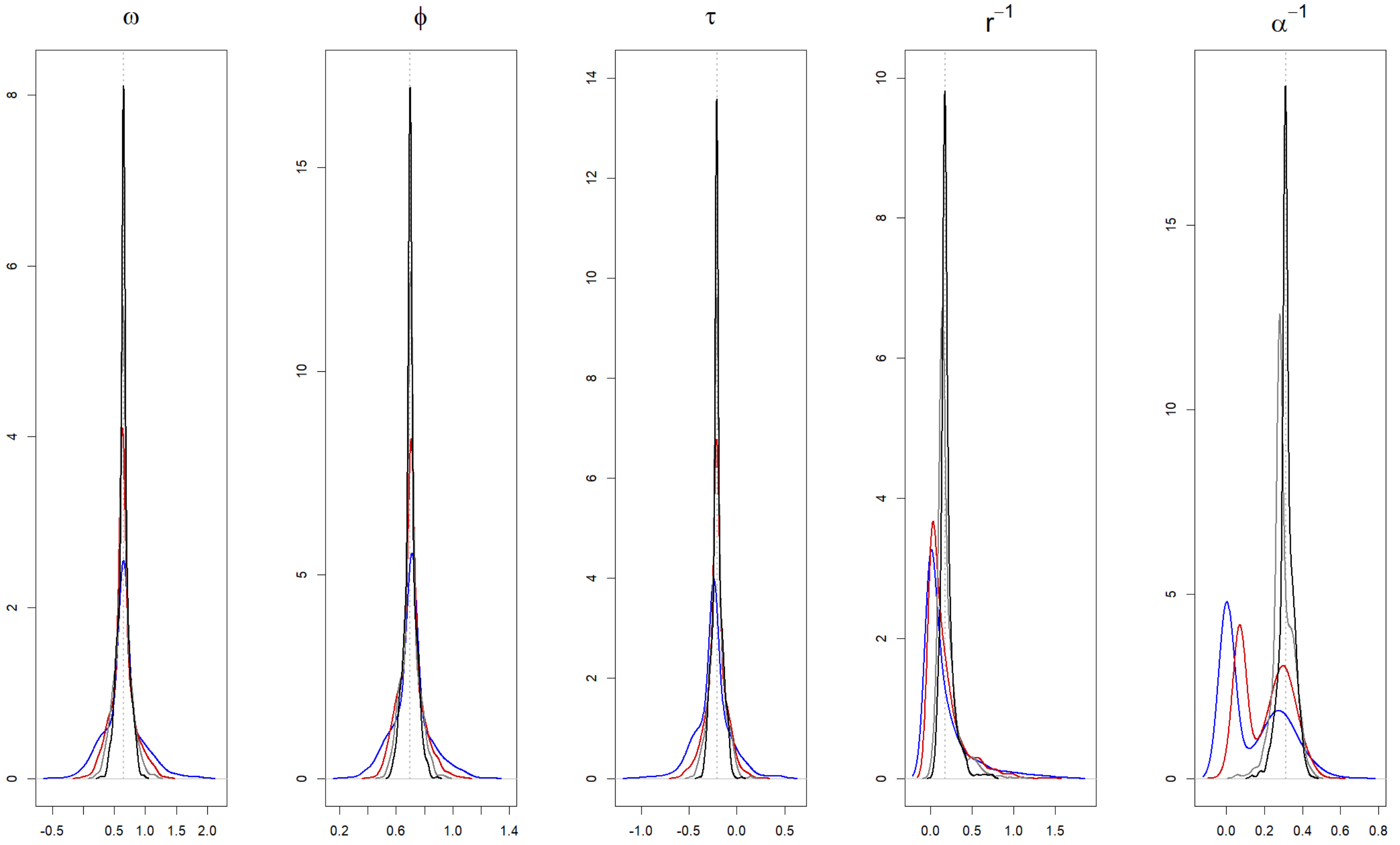

Figure 2.1 Kernel density of the ML estimates obtained from 1000 Monte Carlo replications for sample sizes 100 (—), 250 (—), 500 (—) and 1000 (—). The vertical dotted lines represent true parameter values and the true parameter vector is $(\omega, \phi, \tau, r, \alpha)^T = (0.65, 0.7, -0.2, 5, 3)^T$

Parameter	True Value	Ave Mean	Ave Std		True Value	Ave Mean	Ave Std
$n = 100$				$n = 250$			
ω	0.65	0.6407	0.2976	ω	0.65	0.6412	0.1846
ϕ	0.7	0.6945	0.1286	ϕ	0.7	0.6914	0.0874
τ	-0.2	-0.2392	0.1830	τ	-0.2	-0.2107	0.1075
r^{-1}	0.2	0.1692	0.2740	r^{-1}	0.2	0.1613	0.2001
α^{-1}	0.3333	0.1454	0.1552	α^{-1}	0.3333	0.1992	0.1379

Table 2.2 Simulation results for the ML estimators of the log-linear BNB-INGARCH model obtained from 1000 Monte Carlo replications. The parameters r and α are reparameterized in terms of their inverse

2.5 Empirical application on futures tick count data

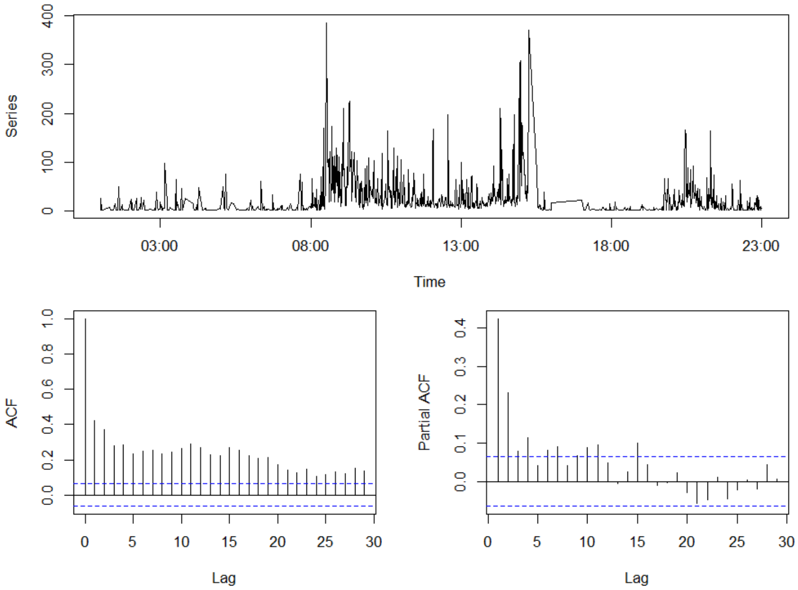

Figure 2.2 Numbers of minute-bar VIX futures tick count data on 02nd January, 2020 from 01:00 a.m. to 11:00 p.m. (top panel) and sample auto-correlation function (bottom left) and partial auto-correlation function of the series (bottom right)

This section illustrates the proposed methodology by an empirical application on the numbers of minute-bar VIX futures tick count data. This historical intraday market data is delivered by Tick Data provider and is available from https://www.tickdata.com/. The sample data set we consider consists of 920 available observations between 01:00 a.m. and 11:00 p.m. on the day January 02, 2020. The empirical mean and variance are 25.789 and 1669.625 respectively, indicating considerable over-dispersion pattern of the data set. Figure 2.2 depicts the plot and the empirical auto-correlation functions of the tick count series. The auto-correlation function (ACF) and partial auto-correlation function (PACF) plots suggest the existence of significant auto-correlations in the data. The series exhibits several extreme observations. Specifically, the number of minute-bar VIX futures tick count is exceedingly high at 08:31 a.m., 14:58 p.m. and 15:15 p.m.. These attributes indicate the desirability of BNB auto-regressions to capture the auto-correlation structure and to account for the extreme observations in the data by means of the heavy-tailedness characteristics of the BNB distribution.

We compare the performances of Bayesian log-linear BNB-INGARCH models and their counterpart log-linear NB-INGARCH models with order specifications $\{(p,q)\} = \{(1,1),(1,2),(2,1),(2,2)\}$. The prior calibrations are presented below. For both BNB-INGARCH and NB-INGARCH models, we adopt normal prior $N(0.5, 0.25^2)$ for ω, constrained normal priors $N(0.2, 0.15^2)$ for ϕ_i and τ_j with $\{\phi_i, \tau_j\}$ satisfying $|\phi_i| < 1, |\tau_j| < 1$, $|\sum_{i=1}^{p} \phi_i + \sum_{j=1}^{q} \tau_j| < 1$ and gamma prior $\mathrm{Ga}(5, 0.5)$ for r. For the tail parameter α under BNB model specifications, we consider the truncated gamma prior $\mathrm{Ga}(5, 0.5)\mathcal{I}(\alpha > 1)$. We also refer to the analogical prior establishments in the simulation analysis section for the considered empirical application and observe the parameter estimations are robust to the selection of priors and hyperparameters. We perform 100,000 MCMC iterations and discard the first 50,000 iterations as a burn-in sample in each model set-up scenario. The remaining 50,000 samples are thinned to every 10th iteration to alleviate autocorrelation, which results in a total of 5000 posterior samples. Table 2.3 summarizes the estimation results including the posterior mean, standard deviation, the posterior 2.5th and 97.5th percentiles, Akaike information criterion (AIC) and Bayesian information criterion (BIC) for each specification separately. The convergence diagnostics of each of the Markov chains is investigated through the trace plots for each parameter of the fitted models. We conclude that proper mixing is achieved by each MCMC sampler. Based on the values of AIC and BIC provided in Table 2.3, the BNB-INGARCH models under all considered order specifications are superior to their corresponding NB-INGARCH counterparts. The preference for the BNB distribution can be further embodied in the low estimate values of the tail parameter α, which is estimated to be around 2.6 with a standard deviation of about 0.2 in most order specification scenarios. This implies the heavy-tailedness of the estimated conditional distribution of the data with only a finite second order moment. Moreover, the BNB-INGARCH(1,2) is found to be a competitive model for this data set among all the considered candidates. Due to space limitation, we only provide the diagnostic trace plots of the Markov chains for each parameter in the favoured BNB-INGARCH(1,2) model in Figure 2.3. Figure 2.4 presents a map of the original tick count time series (in black) and posterior mean predictions under the BNB-INGARCH(1,2) model (depicted by the red line) and the NB-INGARCH(1,2) model (depicted by the blue line). In general, the estimated means closely mimic the count series and capture well the trends within the time series, suggesting the proposed models aptly accommodate the underlying data. Notably, during periods of heightened volatility characterized by elevated counts around 08:30 a.m. and 15:00 p.m., the BNB model outperforms its NB counterpart in precisely describing these extreme observations as tail events and yielding more accurate PMF predictions. It is pertinent to highlight that the BNB models, with their inherent ability to accurately capture pronounced evidence of heavy-tailedness, contribute to an enhanced overall model fit.

To check the adequacy of the best fitted log-linear BNB-INGARCH(1,2) model based on the lowest AIC and BIC values, we examine the standardized Pearson residuals $z_t = \frac{y_t - \mathrm{E}(y_t|\mathcal{F}_{t-1})}{\sqrt{\mathrm{Var}(y_t|\mathcal{F}_{t-1})}}$ proposed by Jung et al. (2006). For correctly specified model, the Pearson residuals should have mean zero and variance one, with no significant auto-correlations. Figure 2.5 presents the series plot and auto-correlation functions of the standardized residuals for the fitted BNB-INGARCH(1,2) model. The plots indicate no significant serial correlations in the residuals. On the basis of the diagnostic checking figures, the auto-correlation characteristics in the futures tick count data can be captured and described by the dynamics of log-linear BNB-INGARCH models. We conclude that the proposed and fitted model is adequate.

Par.	BNB-INGARCH(1,1)				NB-INGARCH(1,1)			
	Mean	Std.	2.50%	97.50%	Mean	Std.	2.50%	97.50%
ω	0.1388	0.0336	0.0859	0.2071	0.2977	0.0713	0.1784	0.4522
ϕ	0.2018	0.0241	0.1535	0.2462	0.2535	0.0349	0.1888	0.3252
τ	0.7780	0.0278	0.7195	0.8337	0.6789	0.0493	0.5750	0.7677
r	5.4710	0.6199	4.6494	6.9842	1.0310	0.0484	0.9384	1.1288
α	2.5587	0.1417	2.2907	2.8484				
AIC	7084.05				7328.54			
BIC	7108.17				7347.84			
Par.	BNB-INGARCH(1,2)				NB-INGARCH(1,2)			
	Mean	Std.	2.50%	97.50%	Mean	Std.	2.50%	97.50%
ω	0.1483	0.0340	0.0892	0.2218	0.2450	0.0534	0.1519	0.3624
ϕ	0.2360	0.0256	0.1884	0.2888	0.2366	0.0297	0.1811	0.2979
τ_1	0.4231	0.0824	0.2672	0.5847	0.4474	0.0874	0.2787	0.6263
τ_2	0.3220	0.0747	0.1754	0.4641	0.2637	0.0814	0.1046	0.4214
r	4.9836	0.8121	3.6044	6.8434	1.0340	0.0475	0.9453	1.1291
α	2.5639	0.1559	2.2722	2.8801				
AIC	7069.43				7304.76			
BIC	7098.38				7328.88			
Par.	BNB-INGARCH(2,1)				NB-INGARCH(2,1)			
	Mean	Std.	2.50%	97.50%	Mean	Std.	2.50%	97.50%
ω	0.1274	0.0376	0.0690	0.2137	0.3395	0.0959	0.1814	0.5520
ϕ_1	0.2176	0.0316	0.1540	0.2797	0.2174	0.0378	0.1452	0.2920
ϕ_2	-0.0202	0.0412	-0.0976	0.0651	0.0823	0.0546	-0.0226	0.1909
τ	0.7855	0.0364	0.7048	0.8515	0.6246	0.0779	0.4615	0.7624
r	4.8684	0.8512	3.3995	6.8476	1.0280	0.0478	0.9369	1.1254
α	2.5717	0.1709	2.2644	2.9241				
AIC	7078.84				7311.16			
BIC	7107.79				7335.28			
Par.	BNB-INGARCH(2,2)				NB-INGARCH(2,2)			
	Mean	Std.	2.50%	97.50%	Mean	Std.	2.50%	97.50%
ω	0.1796	0.0482	0.1040	0.2859	0.3414	0.0817	0.2023	0.5149
ϕ_1	0.2111	0.0299	0.1532	0.2700	0.1976	0.0364	0.1274	0.2701
ϕ_2	0.0723	0.0451	-0.0128	0.1602	0.1338	0.0484	0.0361	0.2273
τ_1	0.2999	0.1159	0.0608	0.5256	0.2425	0.1189	0.0115	0.4786
τ_2	0.3935	0.0907	0.2172	0.5762	0.3529	0.0938	0.1701	0.5371
r	4.9360	0.7878	3.5750	6.6344	1.0345	0.0466	0.9463	1.1283
α	2.5674	0.1700	2.2490	2.9613				
AIC	7071.68				7304.86			
BIC	7105.45				7333.80			

Table 2.3 Summary of estimation results for VIX futures tick count data

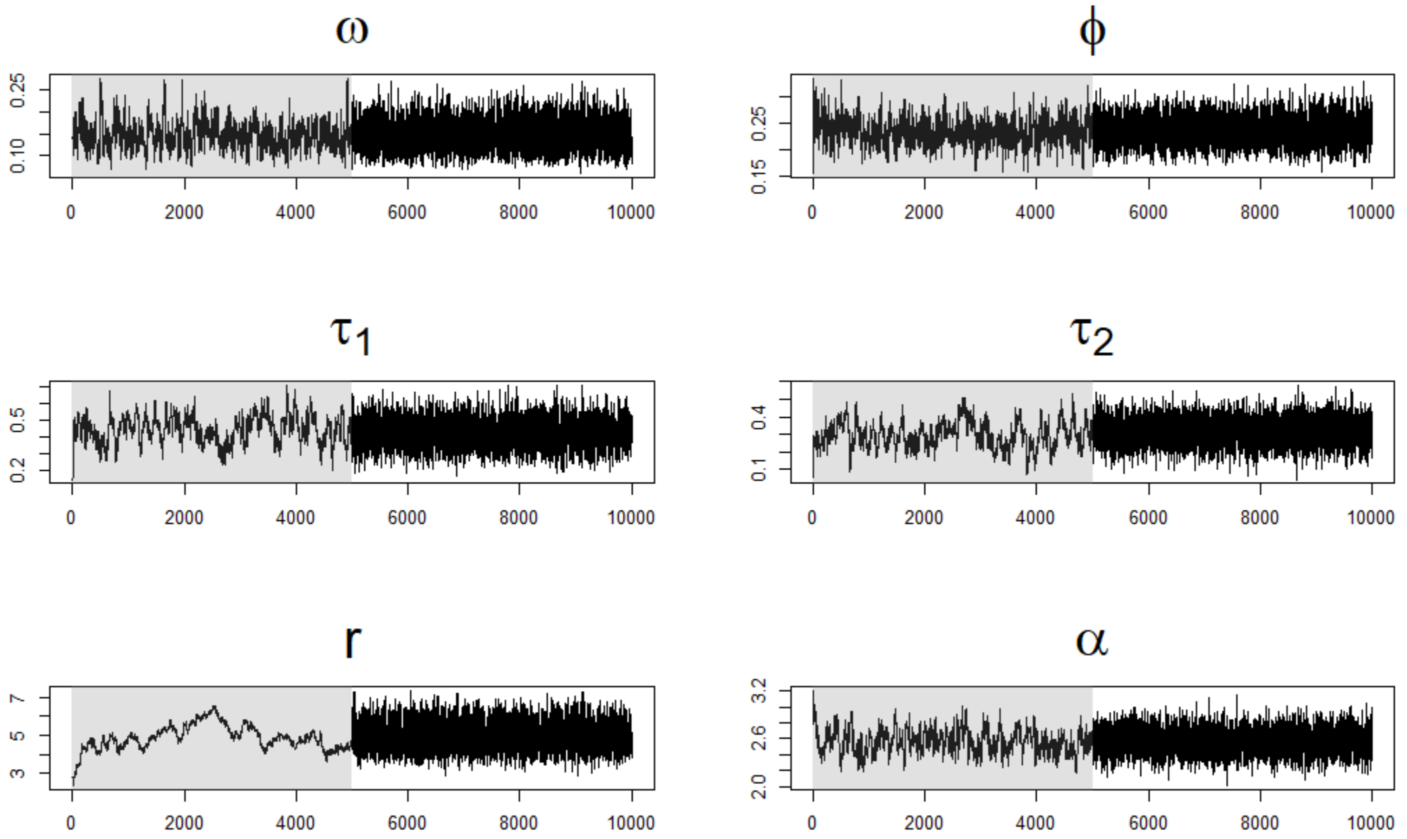

Figure 2.3 Trace plots of the Markov chains for the log-linear BNB-INGARCH(1,2) model marginal posterior distributions of parameters $\omega, \phi, \tau_1, \tau_2, r$ and α for VIX futures tick count data. The grey rectangles represent the burn-in periods

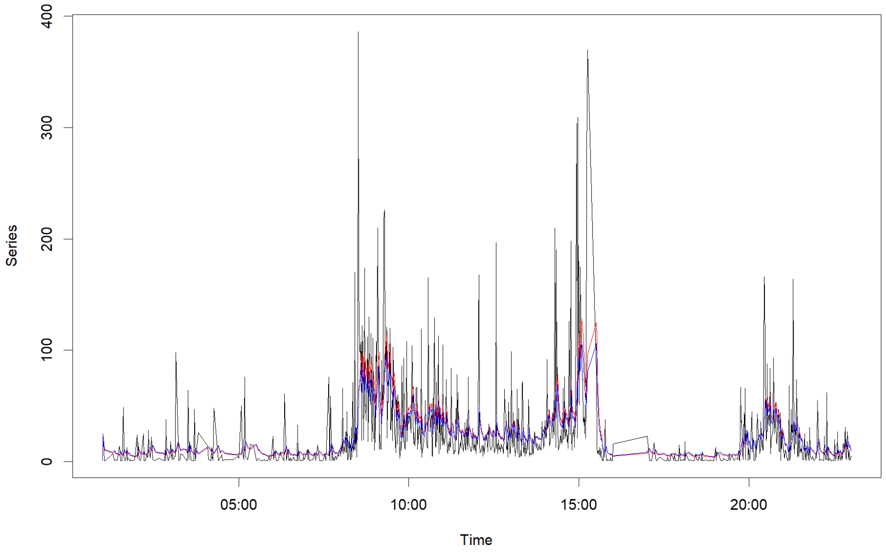

Figure 2.4 Estimated posterior means of the log-linear BNB-INGARCH(1,2) model (in red) and the log-linear NB-INGARCH(1,2) model (in blue). The original tick count series is depicted in black

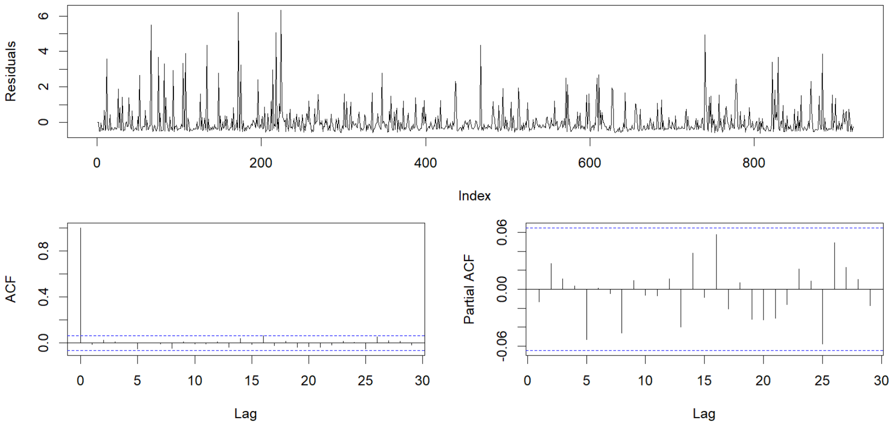

Figure 2.5 The top panel reports the standardized Pearson residuals of the log-linear BNB-INGARCH(1,2) model for VIX futures tick count data. The bottom left panel and bottom right panel present the auto-correlation and partial auto-correlation functions of the residuals respectively

2.6 Chapter summary

The Bayesian estimation approach and model selections for the proposed log-linear Beta–negative binomial integer-valued GARCH model have been presented. Parameter estimations for the proposed process are performed based on adaptive Markov chain Monte Carlo methods. The empirical application on high-frequency intraday VIX futures tick count data indicates that the proposed model is adequate and provides better performance than the INGARCH model under negative binomial distribution assumptions. This chapter, by extending the log-linear BNB-INGARCH model in Bayesian frameworks, complements another aspect of current literature on modelling discrete time series with heavy-tailedness characteristics.

Chapter 3

Bayesian scale mixtures of normals linear regression and Bayesian quantile regression with big data and variable selection

3.1 Introduction

Quantile regression (QR) estimates various conditional quantiles of a response or dependent random variable, including the median (0.5th quantile). Putting different quantile regressions together provides a more complete description of the underlying conditional distribution of the response than a simple mean regression. This is particularly useful when the conditional distribution is asymmetric or heterogeneous or fat-tailed or truncated. Quantile regression has been widely used in statistics and numerous application areas (Cole and Green, 1992; Koenker and Hallock, 2001; Yu et al., 2003; Briollais and Durrieu, 2014, among others). In the "big data" era for statistical science, the richness of data sources with many complicated data structures and the increase of extreme values and heterogeneity may see quantile regression methods more relevant than mean regression to dig deep into the data and grab information from it. In particular, with advanced power of computer technology, complicated quantile regression-based models could be developed under a Bayesian framework, and Bayesian quantile regression (BQR) has received increasing attention from both theoretical and empirical viewpoints with wide applications (see Bernardi et al., 2015; Wang et al., 2016b; Rodrigues and Fan, 2017; Petrella and Raponi, 2019, among others). So far, several methods have been developed to quantile regression for big data analysis (Wu and Yin, 2015; Yu et al., 2017; Gu et al., 2018; Chen et al., 2019b, among others), but little attention has been paid to such methodology under Bayesian inference paradigm.

In this chapter, we propose a new approach of BQR for big data. This approach has its posterior distribution on the whole data as a joint posterior from M sub-datasets split from the whole data. Section 3.2 and Section 3.3 give details of the normal-inverse-gamma (NIG) expressions of the prior and posterior distributions for Bayesian scale mixtures of normals linear regression and BQR respectively. Section 3.4 presents the posterior predictive distributions. Section 3.5 develops big data based algorithms for Bayesian scale mixtures of normals model and BQR via the introduction of NIG summation operator. Section 3.6 provides big data based algorithms for Bayesian LASSO scale mixtures of normals regression and Bayesian LASSO quantile regression. Section 3.7 demonstrates the proposed algorithms via simulations and a real data analysis.

3.2 Bayesian scale mixtures of normals linear regression for big data

3.2.1 Model and likelihood

Consider the scale mixtures of normals linear model

$$y_i = \boldsymbol{x}_i^T \boldsymbol{\beta} + \sigma \epsilon_i, \ \ i = 1, \dots, n,$$

where $\boldsymbol{x}_i$ is a $k \times 1$ vector of predictors for observation y_i, $\boldsymbol{\beta}$ is a $k \times 1$ unknown vector of regression coefficients, $\epsilon_1, \dots, \epsilon_n$ are *i.i.d.* random variables distributed as scale mixtures of normals. That is, $\epsilon_i \overset{d}{=} \sqrt{\zeta_i} z_i$ where z_i follows a standard normal distribution and ζ_i is an independent random variable with some known probability distribution f_{ζ_i} on $(0, \infty)$. σ is an unknown scaling factor. We aim to model the conditional mean $E[y_i | \boldsymbol{x}_i, \zeta_i]$ under Bayesian estimation paradigm. Our primary interest is in inference of the unknown parameters $\boldsymbol{\beta}$ and σ. More compactly, the scale mixtures of normals linear regression in matrix format is specified as

$$\boldsymbol{Y} = \mathbf{X}\boldsymbol{\beta} + \sigma \boldsymbol{\epsilon}, \tag{3.1}$$

where $\boldsymbol{Y} = (y_1, \dots, y_n)^T$ is an $n \times 1$ response vector, $\boldsymbol{X} = (\boldsymbol{x}_1, \dots, \boldsymbol{x}_n)^T$ is an $n \times k$ predictor matrix and $\boldsymbol{\epsilon} = (\epsilon_1, \dots, \epsilon_n)^T$ is an $n \times 1$ scale mixtures of normals disturbances with a mean vector of zeros and $n \times n$ positive definite covariance matrix $\boldsymbol{\Sigma} = \text{diag}(\zeta_1, \dots, \zeta_n)$. Then the conditional likelihood of $\boldsymbol{Y}$ is given by

$$f(\boldsymbol{Y} | \boldsymbol{X}, \boldsymbol{\beta}, \sigma^2, \boldsymbol{\Sigma}) \propto (\sigma^2)^{-\frac{n}{2}} \exp\{-\frac{1}{2\sigma^2} (\boldsymbol{Y} - \boldsymbol{X}\boldsymbol{\beta})^T \boldsymbol{\Sigma}^{-1} (\boldsymbol{Y} - \boldsymbol{X}\boldsymbol{\beta})\}. \tag{3.2}$$

Consider the formulation

$$(\boldsymbol{Y}-\boldsymbol{X}\boldsymbol{\beta})^T\boldsymbol{\Sigma}^{-1}(\boldsymbol{Y}-\boldsymbol{X}\boldsymbol{\beta}) = (\boldsymbol{Y}-\boldsymbol{X}\hat{\boldsymbol{\beta}})^T\boldsymbol{\Sigma}^{-1}(\boldsymbol{Y}-\boldsymbol{X}\hat{\boldsymbol{\beta}}) + (\boldsymbol{\beta}-\hat{\boldsymbol{\beta}})^T(\boldsymbol{X}^T\boldsymbol{\Sigma}^{-1}\boldsymbol{X})(\boldsymbol{\beta}-\hat{\boldsymbol{\beta}}),$$

where $\hat{\boldsymbol{\beta}} = (\boldsymbol{X}^T\boldsymbol{\Sigma}^{-1}\boldsymbol{X})^{-1}\boldsymbol{X}^T\boldsymbol{\Sigma}^{-1}\boldsymbol{Y}$, we can thus rewrite likelihood (3.2) as

$$\begin{aligned}
f(\boldsymbol{Y}|\boldsymbol{X},\boldsymbol{\beta},\sigma^2,\boldsymbol{\Sigma}) \propto & (\sigma^2)^{-\frac{n-k}{2}}\exp\{-\frac{1}{2\sigma^2}(\boldsymbol{Y}-\boldsymbol{X}\hat{\boldsymbol{\beta}})^T\boldsymbol{\Sigma}^{-1}(\boldsymbol{Y}-\boldsymbol{X}\hat{\boldsymbol{\beta}})\} \\
& (\sigma^2)^{-\frac{k}{2}}\exp\{-\frac{1}{2\sigma^2}(\boldsymbol{\beta}-\hat{\boldsymbol{\beta}})^T(\boldsymbol{X}^T\boldsymbol{\Sigma}^{-1}\boldsymbol{X})(\boldsymbol{\beta}-\hat{\boldsymbol{\beta}})\} \qquad (3.3) \\
& = (\sigma^2)^{-(a+\frac{k}{2}+1)}\exp\{-\frac{1}{\sigma^2}[b+\frac{1}{2}(\boldsymbol{\beta}-\boldsymbol{\mu})^T\boldsymbol{\Lambda}(\boldsymbol{\beta}-\boldsymbol{\mu})]\} \\
& \propto IG(a,b)N_k(\boldsymbol{\mu},\sigma^2\boldsymbol{\Lambda}^{-1}), \qquad (3.4)
\end{aligned}$$

where $IG(a,b)$ denotes the inverse-gamma distribution with shape parameter a and scale parameter b. $N_k(\boldsymbol{\mu},\sigma^2\boldsymbol{\Lambda}^{-1})$ denotes the multivariate normal distribution with mean vector $\boldsymbol{\mu}$ and covariance matrix $\sigma^2\boldsymbol{\Lambda}^{-1}$. The represented likelihood (3.4) is a typical structure of a k-dimensional normal-inverse-gamma distribution $NIG_k(\boldsymbol{\mu},\boldsymbol{\Lambda},a,b)$ in terms of parameters $(\boldsymbol{\beta},\sigma^2)$. Here $\boldsymbol{\mu}=\hat{\boldsymbol{\beta}}, \boldsymbol{\Lambda}=\boldsymbol{X}^T\boldsymbol{\Sigma}^{-1}\boldsymbol{X}, a=\frac{n-k-2}{2}, b=\frac{1}{2}(\boldsymbol{Y}-\boldsymbol{X}\hat{\boldsymbol{\beta}})^T\boldsymbol{\Sigma}^{-1}(\boldsymbol{Y}-\boldsymbol{X}\hat{\boldsymbol{\beta}})$.

3.2.2 NIG expressions of posterior distribution

3.2.2.1 Posterior distribution under non-informative prior

The conjugate non-informative prior $f(\boldsymbol{\beta},\sigma^2)\propto\sigma^{-2}$ suggests a specific case of the NIG distribution which is denoted as $NIG_k(\mathbf{0}_k,\mathbf{0}_{k\times k},-\frac{k}{2},0)$. Under this prior, the posterior distribution $f(\boldsymbol{\beta},\sigma^2|\boldsymbol{Y},\boldsymbol{X},\boldsymbol{\Sigma})$ is given by

$$\begin{aligned}
f(\boldsymbol{\beta},\sigma^2|\boldsymbol{Y},\boldsymbol{X},\boldsymbol{\Sigma}) &= f(\sigma^2|\boldsymbol{Y},\boldsymbol{X},\boldsymbol{\Sigma})f(\boldsymbol{\beta}|\sigma^2,\boldsymbol{Y},\boldsymbol{X},\boldsymbol{\Sigma}) \\
&= IG(\widetilde{a},\widetilde{b})N_k(\widetilde{\boldsymbol{\mu}},\sigma^2\widetilde{\boldsymbol{\Lambda}}^{-1}) \\
&\propto(\sigma^2)^{-(\widetilde{a}+\frac{k}{2}+1)}\exp\{-\frac{1}{\sigma^2}[\widetilde{b}+\frac{1}{2}(\boldsymbol{\beta}-\widetilde{\boldsymbol{\mu}})^T\widetilde{\boldsymbol{\Lambda}}(\boldsymbol{\beta}-\widetilde{\boldsymbol{\mu}})]\}.
\end{aligned}$$

Then we denote the joint posterior distribution of $(\boldsymbol{\beta},\sigma^2)$ as $f(\boldsymbol{\beta},\sigma^2|\boldsymbol{Y},\boldsymbol{X},\boldsymbol{\Sigma}) = NIG_k(\widetilde{\boldsymbol{\mu}},\widetilde{\boldsymbol{\Lambda}},\widetilde{a},\widetilde{b})$. Here $\widetilde{\boldsymbol{\mu}} = (\boldsymbol{X}^T\boldsymbol{\Sigma}^{-1}\boldsymbol{X})^{-1}\boldsymbol{X}^T\boldsymbol{\Sigma}^{-1}\boldsymbol{Y}, \widetilde{\boldsymbol{\Lambda}} = \mathbf{X}^T\boldsymbol{\Sigma}^{-1}\boldsymbol{X}, \widetilde{a} = \frac{n-k}{2}, \widetilde{b} = \frac{1}{2}\boldsymbol{Y}^T\boldsymbol{\Sigma}^{-1}\boldsymbol{Y} - \frac{1}{2}\widetilde{\boldsymbol{\mu}}^T\widetilde{\boldsymbol{\Lambda}}\widetilde{\boldsymbol{\mu}}$.

3.2.2.2 Posterior distribution under informative prior

Consider a form of conjugate informative prior for $(\boldsymbol{\beta},\sigma^2)$:

$$\begin{aligned}
f(\boldsymbol{\beta},\sigma^2) &= f(\sigma^2)f(\boldsymbol{\beta}|\sigma^2) \\
&\propto(\sigma^2)^{-(a_0+1)}\exp\{-\frac{b_0}{\sigma^2}\}(\sigma^2)^{-\frac{k}{2}}\exp\{-\frac{1}{2\sigma^2}(\boldsymbol{\beta}-\boldsymbol{\mu}_0)^T\boldsymbol{\Lambda}_0(\boldsymbol{\beta}-\boldsymbol{\mu}_0)\} \\
&= (\sigma^2)^{-(a_0+\frac{k}{2}+1)}\exp\{-\frac{1}{\sigma^2}[b_0+\frac{1}{2}(\boldsymbol{\beta}-\boldsymbol{\mu}_0)^T\boldsymbol{\Lambda}_0(\boldsymbol{\beta}-\boldsymbol{\mu}_0)]\},
\end{aligned}$$

where $f(\sigma^2)$ is $IG(a_0, b_0)$ with prior values a_0, b_0 and $f(\boldsymbol{\beta}|\sigma^2)$ is $N_k(\boldsymbol{\mu}_0, \sigma^2\boldsymbol{\Lambda}_0^{-1})$ with prior values $\boldsymbol{\mu}_0, \boldsymbol{\Lambda}_0$. We can thus calibrate the joint prior as an NIG distribution $f(\beta, \sigma^2) = NIG_k(\boldsymbol{\mu}_0, \boldsymbol{\Lambda}_0, a_0, b_0)$. Under this prior, the posterior distribution is given by

$$\begin{aligned} f(\boldsymbol{\beta}, \sigma^2|\boldsymbol{Y}, \boldsymbol{X}, \boldsymbol{\Sigma}) &= f(\sigma^2|\boldsymbol{Y}, \boldsymbol{X}, \boldsymbol{\Sigma})f(\boldsymbol{\beta}|\sigma^2, \boldsymbol{Y}, \boldsymbol{X}, \boldsymbol{\Sigma}) \\ &= IG(\bar{a}, \bar{b})N_k(\bar{\boldsymbol{\mu}}, \sigma^2\bar{\boldsymbol{\Lambda}}^{-1}) \\ &\propto (\sigma^2)^{-(\bar{a}+\frac{k}{2}+1)} \exp\{-\frac{1}{\sigma^2}[\bar{b} + \frac{1}{2}(\boldsymbol{\beta} - \bar{\boldsymbol{\mu}})^T\bar{\boldsymbol{\Lambda}}(\boldsymbol{\beta} - \bar{\boldsymbol{\mu}})]\}, \end{aligned}$$

which can be denoted as $f(\boldsymbol{\beta}, \sigma^2|\boldsymbol{Y}, \boldsymbol{X}, \boldsymbol{\Sigma}) = NIG_k(\bar{\boldsymbol{\mu}}, \bar{\boldsymbol{\Lambda}}, \bar{a}, \bar{b})$. Here $\bar{\boldsymbol{\mu}} = (\boldsymbol{\Lambda}_0 + \boldsymbol{X}^T\boldsymbol{\Sigma}^{-1}\boldsymbol{X})^{-1}(\boldsymbol{\Lambda}_0\boldsymbol{\mu}_0 + \boldsymbol{X}^T\boldsymbol{\Sigma}^{-1}\boldsymbol{Y}), \bar{\boldsymbol{\Lambda}} = \boldsymbol{\Lambda}_0 + \boldsymbol{X}^T\boldsymbol{\Sigma}^{-1}\boldsymbol{X}, \bar{a} = a_0 + \frac{n}{2}, \bar{b} = b_0 + \frac{1}{2}\boldsymbol{Y}^T\boldsymbol{\Sigma}^{-1}\boldsymbol{Y} + \frac{1}{2}\boldsymbol{\mu}_0^T\boldsymbol{\Lambda}_0\boldsymbol{\mu}_0 - \frac{1}{2}\bar{\boldsymbol{\mu}}^T\bar{\boldsymbol{\Lambda}}\bar{\boldsymbol{\mu}}$.

3.3 Bayesian quantile regression for big data

3.3.1 Model and likelihood

Let y_i be a continuous response variable and $\boldsymbol{x}_i$ a $k \times 1$ vector of predictors for the ith observation, $i = 1, \ldots, n$. Denote $Q_p(y_i|\boldsymbol{x}_i)$ as the pth $(0 < p < 1)$ quantile regression function of y_i given $\boldsymbol{x}_i$. Suppose that all conditional quantiles $Q_p(y_i|\boldsymbol{x}_i)$ can be modelled as $Q_p(y_i|\boldsymbol{x}_i) = \boldsymbol{x}_i^T\boldsymbol{\beta}_p$, where $\boldsymbol{\beta}_p$ is a $k \times 1$ vector of unknown parameters that depends on quantile p. Then the linear Quantile Regression (QR) model for the pth quantile can be denoted as

$$y_i = \boldsymbol{x}_i^T\boldsymbol{\beta}_p + \epsilon_i, \ \ i = 1, \ldots, n,$$

where ϵ_i is the error term whose distribution is assumed to have zero pth quantile. Following Koenker and Bassett (1978b), the estimation for $\boldsymbol{\beta}_p$ proceeds by minimizing

$$\sum_{i=1}^{n} \rho_p(y_i - \boldsymbol{x}_i^T\boldsymbol{\beta}_p), \tag{3.5}$$

where $\rho_p(u) = u\{p - \mathcal{I}(u < 0)\}$ is the check function and $\mathcal{I}(\cdot)$ denotes the indicator function. Equivalently, we can express $\rho_p(u)$ as

$$\rho_p(u) = \frac{|u| + (2p-1)u}{2}. \tag{3.6}$$

According to Yu and Moyeed (2001) and Yu and Stander (2007), minimizing (3.5) is equivalent to maximizing a likelihood function that is based on the asymmetric Laplace distribution (ALD) at specific value of p. Assuming an ALD-based working model such that $\epsilon_i \sim ALD(\kappa, \sigma, p)$ with location parameter $\kappa = 0$, scale parameter $\sigma \in (0, \infty)$ and skewness parameter $p \in (0, 1)$, then the probability density function of ϵ_i is given by

$$f(\epsilon_i; \kappa = 0, \sigma, p) = \frac{p(1-p)}{\sigma} \exp\{-\frac{\rho_p(\epsilon_i)}{\sigma}\}, \ \ i = 1, \ldots, n,$$

where $\rho_p(u)$ is defined in (3.6). Following Reed and Yu (2009) and Kozumi and Kobayashi (2011b), we can represent ϵ_i as a scale mixture of normals with an exponential mixing density as follows:

$$\epsilon_i|v_i, \sigma \sim N((1-2p)v_i, 2\sigma v_i), \;\; v_i|\sigma \sim \text{Exp}(\sigma^{-1}p(1-p)),$$

where $\text{Exp}(\theta)$ denotes an exponential distribution with rate parameter θ. Consequently, the conditional distribution of y_i is normal with mean $\boldsymbol{x}_i^T\boldsymbol{\beta}_p+(1-2p)v_i$ and variance $2\sigma v_i$:

$$y_i|\boldsymbol{\beta}_p, \sigma, v_i, \boldsymbol{x}_i \sim N(\boldsymbol{x}_i^T\boldsymbol{\beta}_p + (1-2p)v_i, 2\sigma v_i), \;\; i = 1, \ldots, n. \tag{3.7}$$

The matrix form of (3.7) is as follows:

$$\boldsymbol{Y}|\boldsymbol{\beta}_p, \sigma, \boldsymbol{v}, \boldsymbol{X}, \boldsymbol{V} \sim N_n(\boldsymbol{X}\boldsymbol{\beta}_p + (1-2p)\boldsymbol{v}, 2\sigma\boldsymbol{V}),$$

where $\boldsymbol{Y} = (y_1, \ldots, y_n)^T$ is an $n \times 1$ response vector, $\boldsymbol{X}$ is an $n \times k$ predictor matrix with ith row $\boldsymbol{x}_i^T$, $\boldsymbol{v} = (v_1, \ldots, v_n)^T$ and $\boldsymbol{V} = \text{diag}(\boldsymbol{v})$. Thus, the conditional likelihood of $\boldsymbol{Y}$ is given by

$$f(\boldsymbol{Y}|\boldsymbol{\beta}_p, \sigma, \boldsymbol{v}, \boldsymbol{X}, \boldsymbol{V}) \propto \sigma^{-n/2} \exp\{-\frac{1}{2\sigma}[\boldsymbol{Y}-\boldsymbol{X}\boldsymbol{\beta}_p-(1-2p)\boldsymbol{v}]^T\frac{\boldsymbol{V}^{-1}}{2}[\boldsymbol{Y}-\boldsymbol{X}\boldsymbol{\beta}_p-(1-2p)\boldsymbol{v}]\}.$$

Let $\boldsymbol{Y}_p^* = \frac{1}{\sqrt{2}}(\boldsymbol{Y} - (1-2p)\boldsymbol{v})$ and $\boldsymbol{X}^* = \frac{1}{\sqrt{2}}\boldsymbol{X}$, then $\boldsymbol{Y}_p^*$ follows a normal-type of conditional likelihood as

$$f(\boldsymbol{Y}_p^*|\boldsymbol{\beta}_p, \sigma, \boldsymbol{X}^*, \boldsymbol{V}) \propto \sigma^{-n/2} \exp\{-\frac{1}{2\sigma}[\boldsymbol{Y}_p^* - \boldsymbol{X}^*\boldsymbol{\beta}_p]^T\boldsymbol{V}^{-1}[\boldsymbol{Y}_p^* - \boldsymbol{X}^*\boldsymbol{\beta}_p]\}. \tag{3.8}$$

Denote further $\hat{\boldsymbol{\beta}}_p = (\boldsymbol{X}^{*T}\boldsymbol{V}^{-1}\boldsymbol{X}^*)^{-1}\boldsymbol{X}^{*T}\boldsymbol{V}^{-1}\boldsymbol{Y}_p^*$, we can rewrite (3.8) as

$$\begin{aligned}
f(\boldsymbol{Y}_p^*|\boldsymbol{\beta}_p, \sigma, \boldsymbol{X}^*, \boldsymbol{V}) \propto \sigma^{-\frac{n-k}{2}} & \exp\{-\frac{1}{2\sigma}[\boldsymbol{Y}_p^* - \boldsymbol{X}^*\hat{\boldsymbol{\beta}}_p]^T\boldsymbol{V}^{-1}[\boldsymbol{Y}_p^* - \boldsymbol{X}^*\hat{\boldsymbol{\beta}}_p]\} \\
& \sigma^{-\frac{k}{2}} \exp\{-\frac{1}{2\sigma}(\boldsymbol{\beta}_p - \hat{\boldsymbol{\beta}}_p)^T(\boldsymbol{X}^{*T}\boldsymbol{V}^{-1}\boldsymbol{X}^*)(\boldsymbol{\beta}_p - \hat{\boldsymbol{\beta}}_p)\} \\
& = (\sigma)^{-(a+\frac{k}{2}+1)} \exp\{-\frac{1}{\sigma}[b_p + \frac{1}{2}(\boldsymbol{\beta}_p - \boldsymbol{\mu}_p)^T\boldsymbol{\Lambda}(\boldsymbol{\beta}_p - \boldsymbol{\mu}_p)]\} \\
& \propto IG(a, b_p)N_k(\boldsymbol{\mu}_p, \sigma\boldsymbol{\Lambda}^{-1}),
\end{aligned} \tag{3.9}$$

where $\boldsymbol{\mu}_p = \hat{\boldsymbol{\beta}}_p, \boldsymbol{\Lambda} = \boldsymbol{X}^{*T}\boldsymbol{V}^{-1}\boldsymbol{X}^*, a = \frac{n-k-2}{2}, b_p = \frac{1}{2}[\boldsymbol{Y}_p^* - \boldsymbol{X}^*\hat{\boldsymbol{\beta}}_p]^T\boldsymbol{V}^{-1}[\boldsymbol{Y}_p^* - \boldsymbol{X}^*\hat{\boldsymbol{\beta}}_p]$. The reformulated likelihood (3.9) is a structure of a k-dimensional distribution $NIG_k(\boldsymbol{\mu}_p, \boldsymbol{\Lambda}, a, b_p)$ in terms of parameters $(\boldsymbol{\beta}_p, \sigma)$.

3.3.2 NIG expressions of posterior distribution

3.3.2.1 Posterior distribution under non-informative prior

The conjugate non-informative prior $f(\boldsymbol{\beta}_p, \sigma) \propto \sigma^{-1}$ suggests a form of $NIG_k(\boldsymbol{0}_k, \boldsymbol{0}_{k\times k}, -\frac{k}{2}, 0)$. Given this prior, the joint conditional posterior distribution $f(\boldsymbol{\beta}_p, \sigma,$

$\boldsymbol{v}|\boldsymbol{Y}_p^*, \boldsymbol{X}^*)$ can be written as

$$\begin{aligned} f(\boldsymbol{\beta}_p, \sigma, \boldsymbol{v}|\boldsymbol{Y}_p^*, \boldsymbol{X}^*) &\propto f(\boldsymbol{Y}_p^*|\boldsymbol{\beta}_p, \sigma, \boldsymbol{v}) f(\boldsymbol{\beta}_p|\sigma, \boldsymbol{v}) f(\boldsymbol{v}|\sigma) f(\sigma) \\ &\propto \sigma^{-(\frac{3n+2}{2})} (\prod_{i=1}^{n} v_i^{-1/2}) \\ &\times \exp\{-\frac{1}{2\sigma}[(\boldsymbol{Y}_p^* - \boldsymbol{X}^*\boldsymbol{\beta}_p)^T \boldsymbol{V}^{-1} (\boldsymbol{Y}_p^* - \boldsymbol{X}^*\boldsymbol{\beta}_p) + 2p(1-p)\sum_{i=1}^{n} v_i]\}. \end{aligned}$$

The posterior distribution $f(\boldsymbol{\beta}_p, \sigma|\boldsymbol{v}, \boldsymbol{Y}_p^*, \boldsymbol{X}^*)$ is thus given by

$$\begin{aligned} f(\boldsymbol{\beta}_p, \sigma|\boldsymbol{v}, \boldsymbol{Y}_p^*, \boldsymbol{X}^*) &\propto \sigma^{-(\frac{3n+2}{2})} \exp\{-\frac{1}{2\sigma}[(\boldsymbol{Y}_p^* - \boldsymbol{X}^*\boldsymbol{\beta}_p)^T \boldsymbol{V}^{-1} (\boldsymbol{Y}_p^* - \boldsymbol{X}^*\boldsymbol{\beta}_p) + 2p(1-p)\sum_{i=1}^{n} v_i]\} \\ &= \sigma^{-(\frac{3n-k}{2} + \frac{k}{2} + 1)} \exp\{-\frac{1}{\sigma}[\widetilde{b}_p + \frac{1}{2}(\boldsymbol{\beta}_p - \widetilde{\boldsymbol{\mu}}_p)^T \widetilde{\boldsymbol{\Lambda}} (\boldsymbol{\beta}_p - \widetilde{\boldsymbol{\mu}}_p)]\}, \end{aligned}$$

which can be denoted as a k-dimensional distribution $NIG_k(\widetilde{\boldsymbol{\mu}}_p, \widetilde{\boldsymbol{\Lambda}}, \widetilde{a}, \widetilde{b}_p)$, where $\widetilde{\boldsymbol{\mu}}_p = (\boldsymbol{X}^{*T}\boldsymbol{V}^{-1}\boldsymbol{X}^*)^{-1}\boldsymbol{X}^{*T}\boldsymbol{V}^{-1}\boldsymbol{Y}_p^*, \widetilde{\boldsymbol{\Lambda}} = \boldsymbol{X}^{*T}\boldsymbol{V}^{-1}\boldsymbol{X}^*, \widetilde{a} = \frac{3n-k}{2}, \widetilde{b}_p = \frac{1}{2}\boldsymbol{Y}_p^{*T}\boldsymbol{V}^{-1} \boldsymbol{Y}_p^* - \frac{1}{2}Y_p^{*T}\boldsymbol{V}^{-1}\boldsymbol{X}^*\widetilde{\boldsymbol{\mu}}_p + p(1-p)\sum_{i=1}^{n} v_i$. Furthermore, the full posterior distribution of each v_i conditional on $\boldsymbol{\beta}_p, \sigma$ and raw data $y_i, \boldsymbol{x}_i, i = 1, 2, \ldots, n$ is obtained by

$$\begin{aligned} f(v_i|\boldsymbol{\beta}_p, \sigma, y_i, \boldsymbol{x}_i) &\propto v_i^{-1/2} \exp\{-\frac{1}{4\sigma}[v_i^{-1}(y_i - (1-2p)v_i - \boldsymbol{x}_i^T\boldsymbol{\beta}_p)^2] - \frac{p(1-p)}{\sigma} v_i\} \\ &= v_i^{-1/2} \exp\{-\frac{1}{4\sigma}[v_i^{-1}(y_i - \boldsymbol{x}_i^T\boldsymbol{\beta}_p)^2 + v_i]\} \\ &= v_i^{-1/2} \exp\{-\frac{1}{2}(v_i^{-1}\widetilde{\xi}_i^{\,2} + v_i\widetilde{\zeta}_i^{\,2})\}, \end{aligned}$$

where $\widetilde{\xi}_i^{\,2} = (y_i - \boldsymbol{x}_i^T\boldsymbol{\beta}_p)^2/2\sigma$ and $\widetilde{\zeta}_i^{\,2} = 1/2\sigma$. This conditional posterior can be recognized as a form of generalized inverse Gaussian distribution $GIG(\frac{1}{2}, \widetilde{\xi}_i, \widetilde{\zeta}_i)$. Recall that if $z \sim GIG(\varphi, \eta_1, \eta_2)$, then the probability density function of z is given by

$$f(z|\varphi, \eta_1, \eta_2) = \frac{(\eta_2/\eta_1)^{\varphi}}{2K_{\varphi}(\eta_1\eta_2)} z^{\varphi-1} \exp\{-\frac{1}{2}(z^{-1}\,\eta_1^2 + z\,\eta_2^2)\}, z > 0, -\infty < \varphi < \infty, \eta_1, \eta_2 \geqslant 0,$$

where $K_{\varphi}(\cdot)$ is a modified Bessel function of the third kind (Barndorff-Nielsen and Shephard, 2001).

3.3.2.2 Posterior distribution under informative g-prior

For the informative prior setting, following Alhamzawi and Yu (2013), a conjugate prior for $(\boldsymbol{\beta}_p, \sigma)$ with a modification of Zellner's informative g-prior (Zellner, 1986) in QR could be provided as

$$\boldsymbol{\beta}_p|\sigma, \boldsymbol{X}^*, \boldsymbol{V} \sim N_k(\boldsymbol{0}_k, g\sigma(\boldsymbol{X}^{*T}\boldsymbol{V}^{-1}\boldsymbol{X}^*)^{-1}), \;\; f(\sigma) \propto \sigma^{-1},$$

where $g > 0$ is a known scaling factor prescribed by the user. Smith and Kohn (1996) proposed a Bayesian variable selection algorithm utilizing regression splines. They found that the choice of $g = 100$ works well and suggested to choose g between 10 and 1000. Following Smith and Kohn (1996), the fixed setting of $g = 100$ has been considered by some other authors (see Lee et al., 2003; Gupta et al., 2007, among others). Then we obtain the joint prior distribution of $(\boldsymbol{\beta}_p, \sigma)$ as

$$f(\boldsymbol{\beta}_p, \sigma|\boldsymbol{X}^*, \boldsymbol{V}) \propto \sigma^{-(\frac{k}{2}+1)} \exp\{-\frac{1}{\sigma}[\frac{1}{2}\boldsymbol{\beta}_p^T \frac{\boldsymbol{X}^{*T}\boldsymbol{V}^{-1}\boldsymbol{X}^*}{g}\boldsymbol{\beta}_p]\}, \tag{3.10}$$

which is a special case of $NIG_k(\boldsymbol{\mu}_0, \boldsymbol{\Lambda}_{g0}, a_0, b_0)$ with $\boldsymbol{\mu}_0 = \boldsymbol{0}_k, \boldsymbol{\Lambda}_{g0} = \frac{\boldsymbol{X}^{*T}\boldsymbol{V}^{-1}\boldsymbol{X}^*}{g}$, $a_0 = 0, b_0 = 0$.

The joint conditional posterior distribution $f(\boldsymbol{\beta}_p, \sigma, \boldsymbol{v}|\boldsymbol{Y}_p^*, \boldsymbol{X}^*)$ under prior (3.10) is given by

$$f(\boldsymbol{\beta}_p, \sigma, \boldsymbol{v}|\boldsymbol{Y}_p^*, \boldsymbol{X}^*) \propto f(\boldsymbol{Y}_p^*|\boldsymbol{\beta}_p, \sigma, \boldsymbol{v}) f(\boldsymbol{\beta}_p|\sigma, \boldsymbol{v}) f(\boldsymbol{v}|\sigma) f(\sigma)$$
$$\propto \sigma^{-(\frac{3n+k+2}{2})}(\prod_{i=1}^{n} v_i^{-1/2})|\boldsymbol{X}^{*T}\boldsymbol{V}^{-1}\boldsymbol{X}^*|^{1/2}$$
$$\times \exp\{-\frac{1}{2\sigma}[(\boldsymbol{Y}_p^* - \boldsymbol{X}^*\boldsymbol{\beta}_p)^T\boldsymbol{V}^{-1}(\boldsymbol{Y}_p^* - \boldsymbol{X}^*\boldsymbol{\beta}_p) + \boldsymbol{\beta}_p^T\frac{\boldsymbol{X}^{*T}\boldsymbol{V}^{-1}\boldsymbol{X}^*}{g}\boldsymbol{\beta}_p + 2p(1-p)\sum_{i=1}^{n} v_i]\}.$$

The corresponding posterior $f(\boldsymbol{\beta}_p, \sigma|\boldsymbol{v}, \boldsymbol{Y}_p^*, \boldsymbol{X}^*)$ is given as follows:

$$\begin{aligned} f(\boldsymbol{\beta}_p, \sigma|\boldsymbol{v}, \boldsymbol{Y}_p^*, \boldsymbol{X}^*) \propto \sigma^{-(\frac{3n+k+2}{2})} & \exp\{-\frac{1}{2\sigma}[(\boldsymbol{Y}_p^* - \boldsymbol{X}^*\boldsymbol{\beta}_p)^T\boldsymbol{V}^{-1}(\boldsymbol{Y}_p^* - \boldsymbol{X}^*\boldsymbol{\beta}_p) \\ & + \boldsymbol{\beta}_p^T\frac{\boldsymbol{X}^{*T}\boldsymbol{V}^{-1}\boldsymbol{X}^*}{g}\boldsymbol{\beta}_p + 2p(1-p)\sum_{i=1}^{n} v_i]\} \\ & = \sigma^{-(\frac{3n}{2}+\frac{k}{2}+1)} \exp\{-\frac{1}{\sigma}[\bar{b}_p + \frac{1}{2}(\boldsymbol{\beta}_p - \bar{\boldsymbol{\mu}}_p)^T\bar{\boldsymbol{\Lambda}}(\boldsymbol{\beta}_p - \bar{\boldsymbol{\mu}}_p)]\}, \end{aligned}$$

which has an expression of $NIG_k(\bar{\boldsymbol{\mu}}_p, \bar{\boldsymbol{\Lambda}}, \bar{a}, \bar{b}_p)$, where $\bar{\boldsymbol{\mu}}_p = [(1+\frac{1}{g})\boldsymbol{X}^{*T}\boldsymbol{V}^{-1}\boldsymbol{X}^*]^{-1}$ $\boldsymbol{X}^{*T}\boldsymbol{V}^{-1}\boldsymbol{Y}_p^*, \bar{\boldsymbol{\Lambda}} = (1+\frac{1}{g})\boldsymbol{X}^{*T}\boldsymbol{V}^{-1}\boldsymbol{X}^*, \bar{a} = \frac{3n}{2}, \bar{b}_p = \frac{1}{2}\boldsymbol{Y}_p^{*T}\boldsymbol{V}^{-1}\boldsymbol{Y}_p^* - \frac{1}{2}\bar{\boldsymbol{\mu}}_p^T\bar{\boldsymbol{\Lambda}}\bar{\boldsymbol{\mu}}_p +$ $p(1-p)\sum_{i=1}^{n} v_i$. Moreover, the full conditional marginal distributions of $\boldsymbol{\beta}_p$ and σ can be obtained respectively by

$$f(\boldsymbol{\beta}_p|\sigma, \boldsymbol{v}, \boldsymbol{Y}_p^*, \boldsymbol{X}^*) \propto \exp\{-\frac{1}{2\sigma}[(\boldsymbol{Y}_p^* - \boldsymbol{X}^*\boldsymbol{\beta}_p)^T\boldsymbol{V}^{-1}(\boldsymbol{Y}_p^* - \boldsymbol{X}^*\boldsymbol{\beta}_p) + \boldsymbol{\beta}_p^T\frac{\boldsymbol{X}^{*T}\boldsymbol{V}^{-1}\boldsymbol{X}^*}{g}\boldsymbol{\beta}_p]\},$$

which can be expressed as an $N_k(\bar{\boldsymbol{\mu}}_p, \sigma\bar{\boldsymbol{\Lambda}}^{-1})$, and

$$\begin{aligned} f(\sigma|\boldsymbol{\beta}_p, \boldsymbol{v}, \boldsymbol{Y}_p^*, \boldsymbol{X}^*) \propto \sigma^{-(\frac{3n+k}{2}+1)} & \exp\{-\frac{1}{2\sigma}[(\boldsymbol{Y}_p^* - \boldsymbol{X}^*\boldsymbol{\beta}_p)^T\boldsymbol{V}^{-1}(\boldsymbol{Y}_p^* - \boldsymbol{X}^*\boldsymbol{\beta}_p) \\ & + \boldsymbol{\beta}_p^T\frac{\boldsymbol{X}^{*T}\boldsymbol{V}^{-1}\boldsymbol{X}^*}{g}\boldsymbol{\beta}_p + 2p(1-p)\sum_{i=1}^{n} v_i]\}, \end{aligned}$$

which is an IG distribution with shape $\frac{3n+k}{2}$ and scale $\frac{1}{2}[(\boldsymbol{Y}_p^* - \boldsymbol{X}^*\boldsymbol{\beta}_p)^T \boldsymbol{V}^{-1}(\boldsymbol{Y}_p^* - \boldsymbol{X}^*\boldsymbol{\beta}_p) + \boldsymbol{\beta}_p^T \frac{\boldsymbol{X}^{*T}\boldsymbol{V}^{-1}\boldsymbol{X}^*}{g}\boldsymbol{\beta}_p + 2p(1-p)\sum_{i=1}^n v_i]$. The full posterior distribution of each $v_i, i = 1, 2, \ldots, n$ is also tractable:

$$
\begin{aligned}
f(v_i|\boldsymbol{\beta}_p, \sigma, y_i, \boldsymbol{x}_i) &\propto v_i^{-1} \exp\{-\frac{1}{4\sigma}[v_i^{-1}((y_i - (1-2p)v_i - \boldsymbol{x}_i^T\boldsymbol{\beta}_p)^2 + \frac{\boldsymbol{\beta}_p^T\boldsymbol{x}_i\boldsymbol{x}_i^T\boldsymbol{\beta}_p}{g})] - \frac{p(1-p)}{\sigma}v_i\} \\
&= v_i^{-1} \exp\{-\frac{1}{4\sigma}[v_i^{-1}((y_i - \boldsymbol{x}_i^T\boldsymbol{\beta}_p)^2 + \frac{\boldsymbol{\beta}_p^T\boldsymbol{x}_i\boldsymbol{x}_i^T\boldsymbol{\beta}_p}{g}) + v_i]\} \\
&= v_i^{-1} \exp\{-\frac{1}{2}(v_i^{-1}\bar{\xi_i}^2 + v_i\bar{\zeta_i}^2)\},
\end{aligned}
$$

where $\bar{\xi_i}^2 = [(y_i - \boldsymbol{x}_i^T\boldsymbol{\beta}_p)^2 + \boldsymbol{\beta}_p^T\boldsymbol{x}_i\boldsymbol{x}_i^T\boldsymbol{\beta}_p/g]/2\sigma$ and $\bar{\zeta_i}^2 = 1/2\sigma$, which can be recognized as a $GIG(0, \bar{\xi}_i, \bar{\zeta}_i)$.

3.4 Posterior predictive distributions

3.4.1 Posterior predictive distribution for Bayesian scale mixtures of normals regression

Given a new $n \times k$ predictor matrix $\boldsymbol{X}^{\text{new}}$, one may be interested in the Bayesian prediction of a new response outcome $\boldsymbol{Y}^{\text{new}}$ under the current posterior calibration of $(\boldsymbol{\beta}, \sigma^2)$ with the observations $\boldsymbol{X}, \boldsymbol{Y}$. To obtain the analytic expression of $f(\boldsymbol{Y}^{\text{new}}|\boldsymbol{Y})$, we first derive the following computation result of integrating out σ^2 from the joint posterior $f(\boldsymbol{\beta}, \sigma^2|\boldsymbol{Y}) = NIG_k(\bar{\boldsymbol{\mu}}, \bar{\boldsymbol{\Lambda}}, \bar{a}, \bar{b})$, where the expressions for $\bar{\boldsymbol{\mu}}, \bar{\boldsymbol{\Lambda}}, \bar{a}$ and $\bar{b}$ are given in Section 3.2.2.2.

$$
\begin{aligned}
\int_0^\infty NIG_k(\bar{\boldsymbol{\mu}}, \bar{\boldsymbol{\Lambda}}, \bar{a}, \bar{b})\, d\sigma^2 &= \frac{\bar{b}^{\bar{a}}}{(2\pi)^{\frac{k}{2}}|\bar{\boldsymbol{\Lambda}}^{-1}|^{\frac{1}{2}}\Gamma(\bar{a})} \\
&\quad \int_0^\infty (\sigma^2)^{-(\bar{a}+\frac{k}{2}+1)} \exp\{-\frac{1}{\sigma^2}[\bar{b} + \frac{1}{2}(\boldsymbol{\beta} - \bar{\boldsymbol{\mu}})^T\bar{\boldsymbol{\Lambda}}(\boldsymbol{\beta} - \bar{\boldsymbol{\mu}})]\}\, d\sigma^2 \\
&= \frac{\bar{b}^{\bar{a}}\Gamma(\bar{a} + \frac{k}{2})}{(2\pi)^{\frac{k}{2}}|\bar{\boldsymbol{\Lambda}}^{-1}|^{\frac{1}{2}}\Gamma(\bar{a})}[\bar{b} + \frac{1}{2}(\boldsymbol{\beta} - \bar{\boldsymbol{\mu}})^T\bar{\boldsymbol{\Lambda}}(\boldsymbol{\beta} - \bar{\boldsymbol{\mu}})]^{-(\bar{a}+\frac{k}{2})} \\
&= \frac{\Gamma(\frac{2\bar{a}+k}{2})}{\Gamma(\frac{2\bar{a}}{2})(2\bar{a})^{\frac{k}{2}}\pi^{\frac{k}{2}}|\frac{\bar{b}}{\bar{a}}\bar{\boldsymbol{\Lambda}}^{-1}|^{\frac{1}{2}}}[1 + \frac{1}{2\bar{a}}(\boldsymbol{\beta} - \bar{\boldsymbol{\mu}})^T(\frac{\bar{b}}{\bar{a}}\bar{\boldsymbol{\Lambda}}^{-1})^{-1}(\boldsymbol{\beta} - \bar{\boldsymbol{\mu}})]^{-(\frac{2\bar{a}+k}{2})} \\
&= \frac{\Gamma(\frac{v_t+k}{2})}{\Gamma(\frac{v_t}{2})v_t^{\frac{k}{2}}\pi^{\frac{k}{2}}|\boldsymbol{\Sigma}_t|^{\frac{1}{2}}}[1 + \frac{1}{v_t}(\boldsymbol{\beta} - \boldsymbol{\mu}_t)^T\boldsymbol{\Sigma}_t^{-1}(\boldsymbol{\beta} - \boldsymbol{\mu}_t)]^{-\frac{v_t+k}{2}} \\
&= \boldsymbol{t}_{v_t}(\boldsymbol{\mu}_t, \boldsymbol{\Sigma}_t).
\end{aligned}
\tag{3.11}
$$

That is, the marginal posterior $f(\boldsymbol{\beta}|\boldsymbol{Y})$ is a k-dimensional multivariate t-distribution $\boldsymbol{t}_{v_t}(\boldsymbol{\mu}_t, \boldsymbol{\Sigma}_t)$ with location vector $\boldsymbol{\mu}_t = \bar{\boldsymbol{\mu}}$, shape matrix $\boldsymbol{\Sigma}_t = \frac{\bar{b}}{\bar{a}}\bar{\boldsymbol{\Lambda}}^{-1}$ and degrees of freedom $v_t = 2\bar{a}$. Then the computation of the posterior predictive distribution of $\boldsymbol{Y}^{\text{new}}$ can be proceeded as follows:

$$\begin{aligned} f(\boldsymbol{Y}^{\text{new}}|\boldsymbol{Y}) &= \int_0^\infty \int_{-\infty}^\infty f(\boldsymbol{Y}^{\text{new}}|\boldsymbol{\beta},\sigma^2) f(\boldsymbol{\beta},\sigma^2|\boldsymbol{Y})\, d\boldsymbol{\beta}\, d\sigma^2 \\ &= \int_0^\infty \int_{-\infty}^\infty N_n(\boldsymbol{X}^{\text{new}}\boldsymbol{\beta}, \sigma^2\boldsymbol{\Sigma}) NIG_k(\bar{\boldsymbol{\mu}}, \bar{\boldsymbol{\Lambda}}, \bar{a}, \bar{b})\, d\boldsymbol{\beta}\, d\sigma^2 \\ &= \int_0^\infty NIG_k(\boldsymbol{X}^{\text{new}}\bar{\boldsymbol{\mu}}, (\boldsymbol{\Sigma} + \boldsymbol{X}^{\text{new}}\bar{\boldsymbol{\Lambda}}^{-1}\boldsymbol{X}^{\text{new}T})^{-1}, \bar{a}, \bar{b})\, d\sigma^2. \quad (3.12) \end{aligned}$$

Applying the integral result (3.11) to (3.12), the computation of the density $f(\mathbf{Y}^{\text{new}}|\mathbf{Y})$ is given by

$$f(\boldsymbol{Y}^{\text{new}}|\boldsymbol{Y}) = \boldsymbol{t}_{2\bar{a}}(\boldsymbol{X}^{\text{new}}\bar{\boldsymbol{\mu}}, \frac{\bar{b}}{\bar{a}}(\boldsymbol{\Sigma} + \boldsymbol{X}^{\text{new}}\bar{\boldsymbol{\Lambda}}^{-1}\boldsymbol{X}^{\text{new}T})),$$

which is an n-dimensional multivariate t-distribution with location $\boldsymbol{X}^{\text{new}}\bar{\boldsymbol{\mu}}$, shape matrix $\frac{\bar{b}}{\bar{a}}(\boldsymbol{\Sigma} + \boldsymbol{X}^{\text{new}}\bar{\boldsymbol{\Lambda}}^{-1}\boldsymbol{X}^{\text{new}T})$ and degrees of freedom $2\bar{a}$. Furthermore, by the law of total conditional variance (Bowsher and Swain (2012)), we can obtain the variance of the future observation $\boldsymbol{Y}^{\text{new}}$ conditional on σ^2

$$\begin{aligned} var(\boldsymbol{Y}^{\text{new}}|\sigma^2) &= E[var(\boldsymbol{Y}^{\text{new}}|\boldsymbol{\beta},\sigma^2)|\sigma^2] + var[E(\boldsymbol{Y}^{\text{new}}|\boldsymbol{\beta},\sigma^2)|\sigma^2] \\ &= E[\sigma^2\boldsymbol{\Sigma}|\sigma^2] + var[\boldsymbol{X}^{\text{new}}\boldsymbol{\beta}|\sigma^2] \\ &= (\boldsymbol{\Sigma} + \boldsymbol{X}^{\text{new}}\bar{\boldsymbol{\Lambda}}^{-1}\boldsymbol{X}^{\text{new}T})\sigma^2. \end{aligned}$$

Therefore, given σ^2, the posterior predictive distribution has two constituents of uncertainty: (1) the model variability induced by the term σ^2 in $\boldsymbol{Y}$ and (2) the posterior uncertainty within the current calibration of $(\boldsymbol{\beta},\sigma^2)$ due to the finite sample size of $\boldsymbol{Y}$.

3.4.2 Posterior predictive distribution for Bayesian quantile regression

In the context of the Bayesian quantile regression model, we carry out the prediction of a new measurement $\boldsymbol{Y}^{\text{new}}$ given a new predictor matrix $\boldsymbol{X}^{\text{new}}$ along with the current estimated parameters $(\boldsymbol{\beta}_p,\sigma)$ as follows. Consider the linear QR model for the pth quantile and observations $\boldsymbol{X}$ and $\boldsymbol{Y}$, and follow the notations for $\boldsymbol{X}^*, \boldsymbol{Y}_p^*, \boldsymbol{v}$ and $\boldsymbol{V}$ presented in Section 3.3.1. Under the joint posterior $f(\boldsymbol{\beta}_p, \sigma|\boldsymbol{v}, \mathbf{Y}_p^*, \mathbf{X}^*) = NIG_k(\bar{\boldsymbol{\mu}}_p, \bar{\boldsymbol{\Lambda}}, \bar{a}, \bar{b}_p)$, where $\bar{\boldsymbol{\mu}}_p, \bar{\boldsymbol{\Lambda}}, \bar{a}$ and $\bar{b}_p$ are given in Section 3.2.2, we can proceed with the prediction of $\boldsymbol{Y}^{\text{new}}$ in two steps: (1) let $\boldsymbol{X}^{\text{new}*} = \frac{1}{\sqrt{2}}\boldsymbol{X}^{\text{new}}$ and compute the corresponding conditional density $f(\boldsymbol{Y}_p^{\text{new}*}|\boldsymbol{Y}_p^*)$ (with conditioning on $\boldsymbol{X}^{\text{new}*}$ implicit), where $\boldsymbol{Y}_p^{\text{new}*} = \frac{1}{\sqrt{2}}(\boldsymbol{Y}^{\text{new}} - (1-2p)\boldsymbol{v})$ is a linear transformation of variable $\boldsymbol{Y}^{\text{new}}$; (2) derive the target density $f(\boldsymbol{Y}^{\text{new}}|\boldsymbol{Y}_p^*)$. The conditional distribution of $\boldsymbol{Y}_p^{\text{new}*}$

is given by

$$\begin{aligned} f(\boldsymbol{Y}_p^{\text{new}*}|\boldsymbol{Y}_p^*) &= \int_0^\infty \int_{-\infty}^\infty f(\boldsymbol{Y}_p^{\text{new}*}|\boldsymbol{\beta}_p, \sigma) f(\boldsymbol{\beta}_p, \sigma|\boldsymbol{Y}_p^*)\, d\boldsymbol{\beta}_p\, d\sigma \\ &= \int_0^\infty \int_{-\infty}^\infty N_n(\boldsymbol{X}^{\text{new}*}\boldsymbol{\beta}_p, \sigma\boldsymbol{V}) NIG_k(\bar{\boldsymbol{\mu}}_p, \bar{\boldsymbol{\Lambda}}, \bar{a}, \bar{b}_p)\, d\boldsymbol{\beta}_p\, d\sigma \\ &= \int_0^\infty NIG_k(\boldsymbol{X}^{\text{new}*}\bar{\boldsymbol{\mu}}_p, (\boldsymbol{V} + \boldsymbol{X}^{\text{new}*}\bar{\boldsymbol{\Lambda}}^{-1}\boldsymbol{X}^{\text{new}*T})^{-1}, \bar{a}, \bar{b}_p)\, d\sigma \\ &= \mathbf{t}_{2\bar{a}}(\boldsymbol{X}^{\text{new}*}\bar{\boldsymbol{\mu}}_p, \frac{\bar{b}_p}{\bar{a}}(\boldsymbol{V} + \boldsymbol{X}^{\text{new}*}\bar{\boldsymbol{\Lambda}}^{-1}\boldsymbol{X}^{\text{new}*T})). \end{aligned} \tag{3.13}$$

The conditional of $\boldsymbol{Y}^{\text{new}} = \sqrt{2}\boldsymbol{Y}_p^{\text{new}*} + (1-2p)\boldsymbol{v}$ is a linear combination of the deduced distribution (3.13). Following the affine transformation property of the multivariate t-distribution (see Roth, 2012 for more details), the new response outcome $\boldsymbol{Y}^{\text{new}}$ is distributed as

$$f(\boldsymbol{Y}^{\text{new}}|\boldsymbol{Y}_p^*) \;=\; \boldsymbol{t}_{2\bar{a}}(\sqrt{2}\boldsymbol{X}^{\text{new}*}\bar{\boldsymbol{\mu}}_p + (1-2p)\boldsymbol{v}, \frac{2\bar{b}_p}{\bar{a}}(\boldsymbol{V} + \boldsymbol{X}^{\text{new}*}\bar{\boldsymbol{\Lambda}}^{-1}\boldsymbol{X}^{\text{new}*T})), \tag{3.14}$$

which is an n-dimensional multivariate t-distribution with location $\sqrt{2}\boldsymbol{X}^{\text{new}*}\bar{\boldsymbol{\mu}}_p + (1-2p)\boldsymbol{v}$, shape matrix $\frac{2\bar{b}_p}{\bar{a}}(\boldsymbol{V} + \boldsymbol{X}^{\text{new}*}\bar{\boldsymbol{\Lambda}}^{-1}\boldsymbol{X}^{\text{new}*T})$ and degrees of freedom $2\bar{a}$. Accordingly, the posterior predictive distribution sampling for BQR can be achieved as below. For each $l = 1, \ldots, L$, we draw samples $\sigma^{(l)} \sim IG(\bar{a}, \bar{b}_p)$ and $\boldsymbol{\beta}_p^{(l)} \sim N_k(\bar{\boldsymbol{\mu}}_p, \sigma^{(l)}\bar{\boldsymbol{\Lambda}}^{-1})$. The obtained samples $\{\boldsymbol{\beta}_p^{(l)}, \sigma^{(l)}\}_{l=1}^L$ give L replicates from the joint posterior distribution $f(\boldsymbol{\beta}_p, \sigma|\boldsymbol{v}, \boldsymbol{Y}_p^*, \boldsymbol{X}^*) = NIG_k(\bar{\boldsymbol{\mu}}_p, \bar{\boldsymbol{\Lambda}}, \bar{a}, \bar{b}_p)$. For each sample $\{\boldsymbol{\beta}_p^{(l)}, \sigma^{(l)}\}$, we generate $\boldsymbol{Y}_p^{\text{new}*(l)} \sim N_n(\boldsymbol{X}^{\text{new}*}\boldsymbol{\beta}_p^{(l)}, \sigma^{(l)}\boldsymbol{V})$. The resulting $\{\boldsymbol{Y}_p^{\text{new}*(l)}\}_{l=1}^L$ provide draws for the conditional distribution (3.13). Then the corresponding samples $\{\boldsymbol{Y}^{\text{new}(l)}\}_{l=1}^L = \{\sqrt{2}\boldsymbol{Y}_p^{\text{new}*(l)} + (1-2p)\boldsymbol{v}\}_{l=1}^L$ give L replicates from the target posterior predictive density (3.14).

3.5 Big data based algorithms for Bayesian scale mixtures of normals regression and BQR

In this section, we propose two divide-and-conquer algorithms to facilitate the calculation of full data posterior distribution in big data settings for Bayesian scale mixtures of normals regression and BQR respectively. We first introduce the concept of NIG multiplication operator as follows.

3.5.1 NIG multiplication operator of posterior distribution

Given the linear regression model (3.1) with $n \times 1$ response vector $\boldsymbol{Y}$, observed $n \times k$ design matrix $\boldsymbol{X}$ and $n \times n$ positive definite covariance matrix $\boldsymbol{\Sigma}$, where the sample size n is so large that the data cannot be stored on a single computer.

If we partition the big data into M subsets, such that $\boldsymbol{Y} = (\boldsymbol{Y}_1, \ldots, \boldsymbol{Y}_M)^T$, $\boldsymbol{X} = (\boldsymbol{X}_1, \ldots, \boldsymbol{X}_M)^T$ and $\boldsymbol{\Sigma} = \mathrm{diag}(\boldsymbol{\Sigma}_1, \ldots, \boldsymbol{\Sigma}_M)$, where $\boldsymbol{Y}_m$ is an $n_m \times 1$ vector, $\boldsymbol{X}_m$ is an $n_m \times k$ matrix, $\boldsymbol{\Sigma}_m$ is an $n_m \times n_m$ diagonal matrix and $\sum_{m=1}^M n_m = n$, then following (3.3) and given the sub-datasets, the conditional likelihood function (3.2) can be written as

$$\begin{aligned} f(\boldsymbol{Y}|\boldsymbol{X},\boldsymbol{\beta},\sigma^2,\boldsymbol{\Sigma}) &\propto (\sigma^2)^{-(\sum_{m=1}^M n_m - k)/2} \exp\{-\frac{1}{2\sigma^2}\sum_{m=1}^M (\boldsymbol{Y}_m - \boldsymbol{X}_m\hat{\boldsymbol{\beta}})^T \boldsymbol{\Sigma}_m^{-1} (\boldsymbol{Y}_m - \boldsymbol{X}_m\hat{\boldsymbol{\beta}})\} \\ &\times (\sigma^2)^{-\frac{k}{2}} \exp\{-\frac{1}{2\sigma^2}\sum_{m=1}^M (\boldsymbol{\beta} - \hat{\boldsymbol{\beta}})^T (\boldsymbol{X}_m^T \boldsymbol{\Sigma}_m^{-1} \boldsymbol{X}_m)(\boldsymbol{\beta} - \hat{\boldsymbol{\beta}})\}, \end{aligned} \tag{3.15}$$

where $\hat{\boldsymbol{\beta}} = (\sum_{m=1}^M \boldsymbol{X}_m^T \boldsymbol{\Sigma}_m^{-1} \boldsymbol{X}_m)^{-1} \sum_{m=1}^M \boldsymbol{X}_m^T \boldsymbol{\Sigma}_m^{-1} \boldsymbol{Y}_m$. The reformulated expression (3.15) with regard to parameters of interest $(\boldsymbol{\beta}, \sigma^2)$ further indicates a multiplication of M NIG distributions

$$\begin{aligned} f(\boldsymbol{Y}|\boldsymbol{X},\boldsymbol{\beta},\sigma^2,\boldsymbol{\Sigma}) &\propto \prod_{m=1}^M (\sigma^2)^{-(a_m + \frac{k}{2} + 1)} \exp\{-\frac{1}{\sigma^2}[b_m + \frac{1}{2}(\boldsymbol{\beta} - \boldsymbol{\mu}_m)^T \boldsymbol{\Lambda}_m (\boldsymbol{\beta} - \boldsymbol{\mu}_m)]\} \\ &= \prod_{m=1}^M NIG(\boldsymbol{\mu}_m, \boldsymbol{\Lambda}_m, a_m, b_m), \end{aligned}$$

where $\boldsymbol{\mu}_m = \hat{\boldsymbol{\beta}}, \boldsymbol{\Lambda}_m = \boldsymbol{X}_m^T \boldsymbol{\Sigma}_m^{-1} \boldsymbol{X}_m$, $a_m = \frac{n_m - k - 2}{2}$, $b_m = \frac{1}{2}(\boldsymbol{Y}_m - \boldsymbol{X}_m\hat{\boldsymbol{\beta}})^T \boldsymbol{\Sigma}_m^{-1}$ $(\boldsymbol{Y}_m - \boldsymbol{X}_m\hat{\boldsymbol{\beta}})$. Therefore, we have the following **Proposition 3.1**.

Proposition 3.1. *Given regression model (3.1) and the described data partition rule, the whole data based likelihood and all sub-datasets based likelihood functions follow NIG distributions and satisfy*

$$NIG_k(\boldsymbol{\mu}, \boldsymbol{\Lambda}, a, b) = \prod_{m=1}^M NIG(\boldsymbol{\mu}_m, \boldsymbol{\Lambda}_m, a_m, b_m), \tag{3.16}$$

where $\boldsymbol{\mu} = (\sum_{m=1}^M \boldsymbol{\Lambda}_m)^{-1} \sum_{m=1}^M \boldsymbol{\Lambda}_m \boldsymbol{\mu}_m, \boldsymbol{\Lambda} = \sum_{m=1}^M \boldsymbol{\Lambda}_m, a = \sum_{m=1}^M a_m + \frac{(M-1)(k+2)}{2}, b = \sum_{m=1}^M b_m + \frac{1}{2}\sum_{m=1}^M (\boldsymbol{\mu}_m - \boldsymbol{\mu})^T \boldsymbol{\Lambda}_m (\boldsymbol{\mu}_m - \boldsymbol{\mu})$.

Posterior distributions induced by the entire data set can be obtained by combining formulation (3.16) with specific priors imposed on $\boldsymbol{\beta}$ and σ^2. The following **Theorem 3.1** elaborates the acquisition of the posterior density through the use of the *NIG* multiplication operator.

Theorem 3.1. *Suppose the posterior distribution, under the prior* $NIG_k(\boldsymbol{\mu}, \boldsymbol{\Lambda}, a, b)$ *and big data observations* $\boldsymbol{X}, \boldsymbol{Y}$, *be* $NIG_k(\bar{\boldsymbol{\mu}}, \bar{\boldsymbol{\Lambda}}, \bar{a}, \bar{b})$. *Partition the entire data into* M *subsets, then we have the full data posterior distribution*

$$\begin{aligned} f(\boldsymbol{\beta}, \sigma^2|\boldsymbol{Y},\boldsymbol{X},\boldsymbol{\Sigma}) &= NIG_k(\boldsymbol{\mu}, \boldsymbol{\Lambda}, a, b) \prod_{m=1}^M NIG_k(\boldsymbol{\mu}_m, \boldsymbol{\Lambda}_m, a_m, b_m) \\ &= NIG_k(\bar{\boldsymbol{\mu}}, \bar{\boldsymbol{\Lambda}}, \bar{a}, \bar{b}), \end{aligned}$$

where $\bar{\boldsymbol{\mu}} = (\boldsymbol{\Lambda} + \sum_{m=1}^{M} \boldsymbol{X}_m^T \boldsymbol{\Sigma}_m^{-1} \boldsymbol{X}_m)^{-1}(\boldsymbol{\Lambda}\boldsymbol{\mu} + \sum_{m=1}^{M} \boldsymbol{X}_m^T \boldsymbol{\Sigma}_m^{-1} \boldsymbol{Y}_m), \bar{\boldsymbol{\Lambda}} = \boldsymbol{\Lambda} + \sum_{m=1}^{M} \boldsymbol{X}_m^T \boldsymbol{\Sigma}_m^{-1} \boldsymbol{X}_m, \bar{a} = a + \frac{n}{2}, \bar{b} = b + \frac{1}{2}[\sum_{m=1}^{M} \boldsymbol{Y}_m^T \boldsymbol{\Sigma}_m^{-1} \boldsymbol{Y}_m + \boldsymbol{\mu}^T \boldsymbol{\Lambda} \boldsymbol{\mu} - \bar{\boldsymbol{\mu}}^T \bar{\boldsymbol{\Lambda}} \bar{\boldsymbol{\mu}}]$.

Corollary 3.1.1. *The full data posterior distribution under the non-informative prior* $NIG_k(\mathbf{0}_k, \mathbf{0}_{k\times k}, -\frac{k}{2}, 0)$ *can be obtained as* $NIG_k(\widetilde{\boldsymbol{\mu}}, \widetilde{\boldsymbol{\Lambda}}, \widetilde{a}, \widetilde{b})$*, where* $\widetilde{\boldsymbol{\mu}} = (\sum_{m=1}^{M} \boldsymbol{X}_m^T \boldsymbol{\Sigma}_m^{-1} \boldsymbol{X}_m)^{-1} \sum_{m=1}^{M} \boldsymbol{X}_m^T \boldsymbol{\Sigma}_m^{-1} \boldsymbol{Y}_m, \widetilde{\boldsymbol{\Lambda}} = \sum_{m=1}^{M} \boldsymbol{X}_m^T \boldsymbol{\Sigma}_m^{-1} \boldsymbol{X}_m, \widetilde{a} = \frac{n-k}{2}$, $\widetilde{b} = \frac{1}{2}[\sum_{m=1}^{M} \boldsymbol{Y}_m^T \boldsymbol{\Sigma}_m^{-1} \boldsymbol{Y}_m - \widetilde{\boldsymbol{\mu}}^T \widetilde{\boldsymbol{\Lambda}} \widetilde{\boldsymbol{\mu}}]$.

3.5.2 Algorithm for Bayesian scale mixtures of normals regression

The following efficient divide-and-conquer algorithm is provided to facilitate the study of scale mixtures of normals linear regression in big data scenario.

Algorithm 3.1. *Consider the Bayesian scale mixtures of normals linear regression under informative prior* $NIG_k(\boldsymbol{\mu}_0, \boldsymbol{\Lambda}_0, a_0, b_0)$ *for* $(\boldsymbol{\beta}, \sigma^2)$ *and with observed* $n \times k$ *design matrix* $\boldsymbol{X}$*,* $n \times 1$ *response vector* $\boldsymbol{Y}$ *and positive definite* $n \times n$ *diagonal covariance matrix* $\boldsymbol{\Sigma}$*, where the data set is too large to be fit into a single computer. By partitioning the entire data set into* M *subsets and utilizing the aforementioned NIG multiplication operator, we can obtain the full data posterior distribution by the following divide-and-conquer algorithm.*

Step 1 *let* $\boldsymbol{X} = \begin{bmatrix} \boldsymbol{X}_1 \\ \vdots \\ \boldsymbol{X}_M \end{bmatrix}$, $\boldsymbol{Y} = \begin{bmatrix} \mathbf{Y}_1 \\ \vdots \\ \boldsymbol{Y}_M \end{bmatrix}$, $\boldsymbol{\Sigma} = \begin{bmatrix} \boldsymbol{\Sigma}_1 & \cdots & \mathbf{0} \\ \vdots & \ddots & \vdots \\ \mathbf{0} & \cdots & \boldsymbol{\Sigma}_M \end{bmatrix}$*, where* $\boldsymbol{X}_m$ *is an* $n_m \times k$ *predictor matrix,* $\mathbf{Y}_m$ *is an* $n_m \times 1$ *response vector,* $\boldsymbol{\Sigma}_m$ *is an* $n_m \times n_m$ *diagonal covariance matrix,* $m = 1, \ldots, M$ *and* $\sum_{m=1}^{M} n_m = n$.

Step 2 *for each subset, the corresponding likelihood has a representation of* $NIG_k(\boldsymbol{\mu}_m, \boldsymbol{\Lambda}_m, a_m, b_m)$ *distribution for* $(\boldsymbol{\beta}, \sigma^2)$*. Calculate the multiplicative distribution* $NIG_k(\boldsymbol{\mu}, \boldsymbol{\Lambda}, a, b) = \prod_{m=1}^{M} NIG(\boldsymbol{\mu}_m, \boldsymbol{\Lambda}_m, a_m, b_m)$*, then the full data posterior can be acquired by merging the prior* $NIG_k(\boldsymbol{\mu}_0, \boldsymbol{\Lambda}_0, a_0, b_0)$ *with the distribution* $NIG_k(\boldsymbol{\mu}, \boldsymbol{\Lambda}, a, b)$*:*

$$NIG_k(\bar{\boldsymbol{\mu}}, \bar{\boldsymbol{\Lambda}}, \bar{a}, \bar{b}) = NIG_k(\boldsymbol{\mu}_0, \boldsymbol{\Lambda}_0, a_0, b_0) NIG_k(\boldsymbol{\mu}, \boldsymbol{\Lambda}, a, b),$$

where $\bar{\boldsymbol{\mu}} = (\boldsymbol{\Lambda}_0 + \sum_{m=1}^{M} \boldsymbol{X}_m^T \boldsymbol{\Sigma}_m^{-1} \boldsymbol{X}_m)^{-1}(\boldsymbol{\Lambda}_0 \boldsymbol{\mu}_0 + \sum_{m=1}^{M} \boldsymbol{X}_m^T \boldsymbol{\Sigma}_m^{-1} \boldsymbol{Y}_m), \bar{\boldsymbol{\Lambda}} = \boldsymbol{\Lambda}_0 + \sum_{m=1}^{M} \boldsymbol{X}_m^T \boldsymbol{\Sigma}_m^{-1} \boldsymbol{X}_m, \bar{a} = a_0 + \frac{n}{2}, \bar{b} = b_0 + \frac{1}{2}[\sum_{m=1}^{M} \boldsymbol{Y}_m^T \boldsymbol{\Sigma}_m^{-1} \boldsymbol{Y}_m + \boldsymbol{\mu}_0^T \boldsymbol{\Lambda}_0 \boldsymbol{\mu}_0 - \bar{\boldsymbol{\mu}}^T \bar{\boldsymbol{\Lambda}} \bar{\boldsymbol{\mu}}]$.

In the high-dimensional setting ($k \gg n$), the induced multicollinearity of $\boldsymbol{X}$ implies the singularity of $\boldsymbol{X}^T \boldsymbol{\Sigma}^{-1} \boldsymbol{X}$. However, one can always choose proper prior matrix $\boldsymbol{\Lambda}_0$ such that $\boldsymbol{\Lambda}_0 + \sum_{m=1}^{M} \boldsymbol{X}_m^T \boldsymbol{\Sigma}_m^{-1} \boldsymbol{X}_m$ is non-singular and therefore $\bar{\boldsymbol{\mu}}$ is well-defined.

3.5.3 Algorithm for Bayesian quantile regression

Consider the linear QR model for the pth $(0 < p < 1)$ quantile

$$\boldsymbol{Y} = \boldsymbol{X}\boldsymbol{\beta}_p + \boldsymbol{\epsilon}, \tag{3.17}$$

where $\boldsymbol{Y}$ is an $n \times 1$ response vector, $\boldsymbol{X}$ is an $n \times k$ predictor matrix and $\boldsymbol{\epsilon}$ is an $n \times 1$ vector of $ALD(0, \sigma, p)$ disturbances. Following the reformulated conditional likelihood (3.8), model (3.17) is equivalent to

$$\boldsymbol{Y}_p^* = \boldsymbol{X}^*\boldsymbol{\beta}_p + \sqrt{\sigma}\boldsymbol{\epsilon}^*, \tag{3.18}$$

where $\boldsymbol{Y}_p^* = \frac{1}{\sqrt{2}}(\boldsymbol{Y} - (1-2p)\boldsymbol{v})$, $\boldsymbol{X}^* = \frac{1}{\sqrt{2}}\boldsymbol{X}$ and $\boldsymbol{\epsilon}^* \sim N_n(\boldsymbol{0}_n, \boldsymbol{V})$ with $n \times n$ diagonal positive definite covariance matrix $\boldsymbol{V}$. We proceed with Bayesian inference for big data quantile regression through the proposed NIG multiplication operator. We consider model (3.17) under the g-prior (3.10) and partition the entire data into M subsets $(\boldsymbol{X}_m, \boldsymbol{Y}_m)$ with individual sample size $n_m, m = 1, \ldots, M$. Then the posterior distribution for the whole data can be obtained by merging the given prior with the multiplication of M subset NIG distributions induced from the massive observations. Based on this, an efficient divide-and-conquer algorithm for big data BQR is provided as below.

Algorithm 3.2. *Consider a pth $(0 < p < 1)$ Bayesian quantile regression under g-prior (3.10) with the observed $n \times k$ design matrix $\boldsymbol{X}$ and $n \times 1$ response vector $\boldsymbol{Y}$, where the large data cannot be fit into a single computer due to the memory constraint. We obtain the full data posterior distribution by the following divide-and-conquer algorithm.*

Step 1 *partition the entire data into M subsets $\boldsymbol{X}_m, \boldsymbol{Y}_m, m = 1, 2, \ldots, M$, where $\boldsymbol{X}_m$ is an $n_m \times k$ matrix, $\boldsymbol{Y}_m$ is an $n_m \times 1$ vector and $\sum_{m=1}^M n_m = n$.*

Step 2 *for each $\boldsymbol{X}_m, \boldsymbol{Y}_m$, a Gibbs sampler for sampling $\boldsymbol{\beta}_p, \sigma$ and the $n_m \times 1$ latent vector $\boldsymbol{v}_m$ follows the below sub-steps:*

2.1 *denote j as the iteration count. Then set $j = 0$ and establish $(\boldsymbol{\beta}_p^{(j=0)}, \sigma^{(j=0)}, \boldsymbol{v}_m^{(j=0)})$ to some starting values.*

2.2 *let $\boldsymbol{X}_m^* = \frac{1}{\sqrt{2}}\boldsymbol{X}_m$, $\boldsymbol{Y}_{pm}^* = \frac{1}{\sqrt{2}}(\boldsymbol{Y}_m - (1-2p)\boldsymbol{v}_m)$ and $\boldsymbol{V}_m = diag(\boldsymbol{v}_m)$.*

2.3 *follow the full conditional posterior distributions of $\boldsymbol{\beta}_p, \sigma$ and $\boldsymbol{v}_m$ given in Section 3.3.2.2,*

(i) sample $\boldsymbol{v}_m^{(j+1)}$ from its GIG posterior $f(\boldsymbol{v}_m|\boldsymbol{\beta}_p^{(j)}, \sigma^{(j)})$.

(ii) sample $\sigma^{(j+1)}$ from its IG posterior $f(\sigma|\boldsymbol{\beta}_p^{(j)}, \boldsymbol{v}_m^{(j+1)})$.

(iii) sample $\boldsymbol{\beta}_p^{(j+1)}$ from its multivariate normal posterior $f(\boldsymbol{\beta}_p|\sigma^{(j+1)}, \boldsymbol{v}_m^{(j+1)})$.

2.4 *set $j = j + 1$ and return to **Step 2.3** until $j = L$, where L is the number of iteration times.*

Step 3 *calculate the empirical estimates of the means $\bar{\boldsymbol{\beta}}_p$ and $\bar{\sigma}$ separately based on the $(L - B)$ realizations of the Gibbs sequence (discarding the first B iterations as a burn-in). Then generate an n_m i.i.d. sample on $\bar{v}_i$, where $\bar{v}_i \sim GIG(0, \bar{\xi}_i, \bar{\zeta}_i)$ with $\bar{\xi}_i^{\,2} = [(y_i - \boldsymbol{x}_i^T \bar{\boldsymbol{\beta}}_p)^2 + \bar{\boldsymbol{\beta}}_p^T \boldsymbol{x}_i \boldsymbol{x}_i^T \bar{\boldsymbol{\beta}}_p / g]/2\bar{\sigma}$ and $\bar{\zeta}_i^{\,2} = 1/2\bar{\sigma}, i = 1, 2, \ldots, n_m$. Let $\boldsymbol{v}_m^\dagger = (\bar{v}_1, \ldots, \bar{v}_{n_m})^T, \boldsymbol{Y}_{pm}^\dagger = \frac{1}{\sqrt{2}}(\boldsymbol{Y}_m - (1 - 2p)\boldsymbol{v}_m^\dagger)$ and $\boldsymbol{V}_m^\dagger = diag(\boldsymbol{v}_m^\dagger), m = 1, 2, \ldots, M$.*

Step 4 *for each subset, the corresponding likelihood can be represented as a form of $NIG_k(\boldsymbol{\mu}_{pm}, \boldsymbol{\Lambda}_m, a_m, b_{pm})$ distribution for $(\boldsymbol{\beta}_p, \sigma)$. Obtain the multiplicative distribution $NIG_k(\boldsymbol{\mu}_p, \boldsymbol{\Lambda}, a, b_p) = \prod_{m=1}^{M} NIG(\boldsymbol{\mu}_{pm}, \boldsymbol{\Lambda}_m, a_m, b_{pm})$, then the full data posterior is given by merging the g-prior $NIG_k(\boldsymbol{\mu}_0, \boldsymbol{\Lambda}_{g0}, a_0, b_0)$ and the distribution $NIG_k(\boldsymbol{\mu}_p, \boldsymbol{\Lambda}, a, b_p)$:*

$$NIG_k(\bar{\boldsymbol{\mu}}_p, \bar{\boldsymbol{\Lambda}}, \bar{a}, \bar{b}_p) = NIG_k(\boldsymbol{\mu}_0, \boldsymbol{\Lambda}_{g0}, a_0, b_0) NIG_k(\boldsymbol{\mu}_p, \boldsymbol{\Lambda}, a, b_p),$$

where $\bar{\boldsymbol{\mu}}_p = [(1 + \frac{1}{g}) \sum_{m=1}^{M} \boldsymbol{X}_m^{*T} \boldsymbol{V}_m^{\dagger -1} \boldsymbol{X}_m^*)]^{-1} \sum_{m=1}^{M} \boldsymbol{X}_m^{*T} \boldsymbol{V}_m^{\dagger -1} \boldsymbol{Y}_{pm}^\dagger, \bar{\boldsymbol{\Lambda}} = (1 + \frac{1}{g}) \sum_{m=1}^{M} \boldsymbol{X}_m^{*T} \boldsymbol{V}_m^{\dagger -1} \boldsymbol{X}_m^*, \bar{a} = \frac{3n}{2}, \bar{b}_p = \frac{1}{2}[\sum_{m=1}^{M} \boldsymbol{Y}_{pm}^{\dagger * T} \boldsymbol{V}_m^{\dagger -1} \boldsymbol{Y}_{pm}^{\dagger *} - \bar{\boldsymbol{\mu}}_p^T \bar{\boldsymbol{\Lambda}} \bar{\boldsymbol{\mu}}_p] + p(1 - p) \sum_{m=1}^{M} \|\boldsymbol{v}_m^\dagger\|_1$ and $\|\cdot\|_1$ denotes the ℓ_1 norm of a vector.

3.6 Big data based algorithms for variable selection

3.6.1 Algorithm for Bayesian LASSO scale mixtures of normals regression

The LASSO of Tibshirani (1996) was proposed to estimate linear regression coefficients using $L1$-penalized least squares. Consider the linear regression model (3.1), the LASSO shrinkage regression can be formulated as

$$\min_{\boldsymbol{\beta}} (\boldsymbol{Y} - \boldsymbol{X}\boldsymbol{\beta})^T (\boldsymbol{Y} - \boldsymbol{X}\boldsymbol{\beta}) + \sum_{j=1}^{k} \lambda_j |\beta_j|,$$

where $\lambda_1, \ldots, \lambda_k$ are non-negative regularization parameters. In this context, we embrace the modified LASSO criterion introduced by Wang et al. (2007) to impose different tuning parameters λ_j on different regression coefficients, thereby anticipating a larger amount of shrinkage being applied to insignificant coefficients, whereas a more modest degree of shrinkage is expected for significant coefficients. According to Tibshirani (1996), the LASSO estimates can be interpreted as the posterior mode with independent and identical Laplace priors imposed on the regression coefficients. Following Park and Casella (2008), a conditional Laplace prior is given by

$$f(\boldsymbol{\beta}|\sigma^2) = \prod_{j=1}^{k} \frac{\lambda_j}{2\sqrt{\sigma^2}} \exp\{-\lambda_j|\frac{\beta_j}{\sqrt{\sigma^2}}|\}.$$

As suggested in Park and Casella (2008), any inverse-Gamma prior for σ^2 would maintain conjugacy. Here we consider the marginal prior $f(\sigma^2) = IG(a_0, b_0)$, then the joint prior for $f(\beta, \sigma^2)$ is given by

$$f(\boldsymbol{\beta}, \sigma^2) \propto (\sigma^2)^{-(a_0+\frac{k}{2}+1)} \exp\{-b_0\sigma^{-2} - \sum_{j=1}^{k} \lambda_j|\frac{\beta_j}{\sigma}|\}.$$

Given model (3.1), we have the posterior distribution

$$f(\boldsymbol{\beta}, \sigma^2|\boldsymbol{Y}, \boldsymbol{X}, \boldsymbol{\Sigma}) \propto (\sigma^2)^{-(a_0+\frac{n+k}{2}+1)} \exp\{-b_0\sigma^{-2} - \frac{1}{2}\sigma^{-2}(\boldsymbol{Y}-\boldsymbol{X}\boldsymbol{\beta})^T\boldsymbol{\Sigma}^{-1}(\boldsymbol{Y}-\boldsymbol{X}\boldsymbol{\beta}) - \sum_{j=1}^{k} \lambda_j|\frac{\beta_j}{\sigma}|\}.$$

Following the equality given by Andrews and Mallows (1974)

$$\frac{h}{2}\exp\{-h|z|\} = \int_0^\infty \frac{1}{\sqrt{2\pi s}} \exp\{-z^2/(2s)\} \frac{h^2}{2} \exp\{-h^2 s/2\} ds, \; h > 0,$$

and introducing the latent variables $\boldsymbol{\gamma} = (\gamma_1, \ldots, \gamma_k)^T$ with prior $f(\boldsymbol{\gamma}) = \prod_{j=1}^{k} \frac{\lambda_j^2}{2} \exp(-\frac{\lambda_j^2\gamma_j}{2})$, we have the following Bayesian hierarchical model:

$$\begin{aligned}
\boldsymbol{Y}|\boldsymbol{\beta}, \boldsymbol{X}, \boldsymbol{\Sigma} &\sim N_n(\boldsymbol{X}\boldsymbol{\beta}, \sigma^2\boldsymbol{\Sigma}), \\
\boldsymbol{\beta}|\sigma^2, \gamma_1, \ldots, \gamma_k &\sim N_k(\boldsymbol{0}_k, \sigma^2\boldsymbol{D}_\gamma), \\
\boldsymbol{D}_\gamma &= \text{diag}\,(\gamma_1, \ldots, \gamma_k), \\
\sigma^2, \gamma_1, \ldots, \gamma_k &\sim f(\sigma^2)\, d\sigma^2 \prod_{j=1}^{k} \frac{\lambda_j^2}{2} \exp(-\frac{\lambda_j^2\gamma_j}{2})\, d\gamma_j, \\
\sigma^2, \gamma_1, \ldots, \gamma_k &> 0.
\end{aligned}$$

Then we obtain the conditional prior distribution

$$f(\boldsymbol{\beta}, \sigma^2|\boldsymbol{\gamma}) \sim NIG_k(\boldsymbol{0}_k, \boldsymbol{D}_\gamma^{-1}, a_0, b_0), \tag{3.19}$$

where $\boldsymbol{D}_\gamma^{-1} = \text{diag}\,(\gamma_1^{-1}, \ldots, \gamma_k^{-1})$. For the conditional posterior of $\boldsymbol{\gamma}$, we have $\gamma_j^{-1}|\boldsymbol{\beta}, \sigma^2, \boldsymbol{Y}$ following an inverse-Gaussian distribution with parameters $\sqrt{\frac{\lambda_j^2\sigma^2}{\beta_j^2}}$ and λ_j^2 (see Park and Casella, 2008). A corresponding Gibbs sampler algorithm can be provided as below.

Algorithm 3.3. *Consider the Bayesian LASSO scale mixtures of normals regression model with prior specification (3.19). Given the big data $\boldsymbol{X}$ and $\boldsymbol{Y}$, we*

obtain the following Gibbs sampler algorithm.

Step 1 *the same as presented in* **Algorithm** 3.1.

Step 2 *for each subset, the corresponding likelihood has a representation of* $NIG_k(\boldsymbol{\mu}_m, \boldsymbol{\Lambda}_m, a_m, b_m)$ *distribution for* $(\boldsymbol{\beta}, \sigma^2)$. *Calculate the multiplicative distribution* $NIG_k(\boldsymbol{\mu}, \boldsymbol{\Lambda}, a, b) = \prod_{m=1}^{M} NIG(\boldsymbol{\mu}_m, \boldsymbol{\Lambda}_m, a_m, b_m)$, *then iterate the following sub-steps until draws* $(\boldsymbol{\beta}, \sigma^2, \boldsymbol{\gamma})$ *achieve convergence.*

2.1 *given the current draw of* $\boldsymbol{\gamma}$, *compute* $\boldsymbol{D}_\gamma^{-1} = \mathrm{diag}\,(\gamma_1^{-1}, \ldots, \gamma_k^{-1})$; *obtain posterior* $NIG_k(\bar{\boldsymbol{\mu}}, \bar{\boldsymbol{\Lambda}}, \bar{a}, \bar{b}) = NIG_k(\mathbf{0}_k, \boldsymbol{D}_\gamma^{-1}, a_0, b_0)NIG_k(\boldsymbol{\mu}, \boldsymbol{\Lambda}, a, b)$, *where* $\bar{\boldsymbol{\mu}} = (\boldsymbol{D}_\gamma^{-1} + \sum_{m=1}^{M} \boldsymbol{X}_m^T \boldsymbol{\Sigma}_m^{-1} \boldsymbol{X}_m)^{-1} \sum_{m=1}^{M} \boldsymbol{X}_m^T \boldsymbol{\Sigma}_m^{-1} \boldsymbol{Y}_m, \bar{\boldsymbol{\Lambda}} = \boldsymbol{D}_\gamma^{-1} + \sum_{m=1}^{M} \boldsymbol{X}_m^T$ $\boldsymbol{\Sigma}_m^{-1} \boldsymbol{X}_m, \bar{a} = a_0 + \frac{n}{2}, \bar{b} = b_0 + \frac{1}{2}[\sum_{m=1}^{M} \boldsymbol{Y}_m^T \boldsymbol{\Sigma}_m^{-1} \boldsymbol{Y}_m - \bar{\boldsymbol{\mu}}^T \bar{\boldsymbol{\Lambda}} \bar{\boldsymbol{\mu}}]$; *then generate a draw of* $(\boldsymbol{\beta}, \sigma^2)$ *from* $NIG_k(\bar{\boldsymbol{\mu}}, \bar{\boldsymbol{\Lambda}}, \bar{a}, \bar{b})$.

2.2 *given the current draw of* $(\boldsymbol{\beta}, \sigma^2)$, *generate a draw for each* γ_j^{-1} *from the inverse-Gaussian distribution with parameters* $\sqrt{\frac{\lambda_j^2 \sigma^2}{\beta_j^2}}$ *and* $\lambda_j^2, j = 1, 2, \ldots, k$.

In the high-dimensional setting $(k \gg n)$, one can always choose proper prior matrix $\boldsymbol{D}_\gamma^{-1}$ such that $\boldsymbol{D}_\gamma^{-1} + \sum_{m=1}^{M} \boldsymbol{X}_m^T \boldsymbol{\Sigma}_m^{-1} \boldsymbol{X}_m$ is non-singular and therefore $\bar{\boldsymbol{\mu}}$ is well-defined.

3.6.2 Algorithm for Bayesian LASSO quantile regression

Following the notations outlined in Section 3.3.1, the LASSO regularized quantile regression (Li and Zhu, 2008) can be formulated by

$$\min_{\boldsymbol{\beta}_p} \sum_{i=1}^{n} \rho_p(y_i - \boldsymbol{x}_i^T \boldsymbol{\beta}_p) + \lambda \sum_{j=1}^{k} |\beta_{pj}|,$$

where $\boldsymbol{\beta}_p = (\beta_{p1}, \ldots, \beta_{pk})^T$ and $\lambda \geqslant 0$ is a penalization parameter. Consider a conditional Laplace prior

$$f(\boldsymbol{\beta}_p|\sigma) = \prod_{j=1}^{k} \frac{\lambda_j}{2\sqrt{\sigma}} \exp\{-\lambda_j|\frac{\beta_{pj}}{\sqrt{\sigma}}|\},$$

where $\lambda_1, \ldots, \lambda_k$ are non-negative penalization parameters and specify the marginal prior $f(\sigma) = IG(a_0, b_0)$, the prior for $f(\boldsymbol{\beta}_p, \sigma)$ is obtained by

$$f(\boldsymbol{\beta}_p, \sigma) \propto \sigma^{-(a_0 + \frac{k}{2} + 1)} \exp\{-b_0\sigma^{-1} - \sum_{j=1}^{k} \lambda_j |\frac{\beta_{pj}}{\sqrt{\sigma}}|\}.$$

Consider further the reformulated linear QR model (3.18), we have the posterior distribution

$$f(\boldsymbol{\beta}_p, \sigma|\boldsymbol{v}, \boldsymbol{Y}_p^*, \boldsymbol{X}^*) \propto \sigma^{-(a_0+\frac{3n+k}{2}+1)} \exp\{-\sigma^{-1}[b_0 + p\,(1-p)\sum_{i=1}^{n} v_i] \\ - \frac{1}{2}\sigma^{-1}(\boldsymbol{Y}_p^* - \boldsymbol{X}^*\boldsymbol{\beta}_p)^T \boldsymbol{V}^{-1}(\boldsymbol{Y}_p^* - \boldsymbol{X}^*\boldsymbol{\beta}_p) - \sum_{j=1}^{k}\lambda_j|\frac{\beta_{pj}}{\sqrt{\sigma}}|\}.$$

Again, by introducing the latent variables $\boldsymbol{\gamma} = (\gamma_1, \ldots, \gamma_k)^T$ with the prior $f(\boldsymbol{\gamma}) = \prod_{j=1}^{k} \frac{\lambda_j^2}{2}\exp(-\frac{\lambda_j^2\gamma_j}{2})$, we have the following Bayesian hierarchical model:

$$\begin{aligned} \boldsymbol{Y}_p^*|\boldsymbol{\beta}_p, \sigma, \boldsymbol{v}, \boldsymbol{X}^* &\sim N_n(\boldsymbol{X}^*\boldsymbol{\beta}_p, \sigma\boldsymbol{V}), \\ \boldsymbol{\beta}_p|\sigma, \gamma_1, \ldots, \gamma_k &\sim N_k(\boldsymbol{0}_k, \sigma\boldsymbol{D}_\gamma), \\ \boldsymbol{D}_\gamma &= \mathrm{diag}\,(\gamma_1, \ldots, \gamma_k), \\ \sigma, \gamma_1, \ldots, \gamma_k &\sim f(\sigma)\,d\sigma\prod_{j=1}^{k}\frac{\lambda_j^2}{2}\exp(-\frac{\lambda_j^2\gamma_j}{2})\,d\gamma_j, \\ \sigma, \gamma_1, \ldots, \gamma_k &> 0. \end{aligned}$$

Then the conditional prior distribution can be denoted as

$$f(\boldsymbol{\beta}_p, \sigma|\boldsymbol{\gamma}) \sim NIG_k(\boldsymbol{0}_k, \boldsymbol{D}_\gamma^{-1}, a_0, b_0), \tag{3.20}$$

where $\boldsymbol{D}_\gamma^{-1} = \mathrm{diag}\,(\gamma_1^{-1}, \ldots, \gamma_k^{-1})$. For the conditional posterior of γ_j, we have $\gamma_j^{-1}|\boldsymbol{\beta}_p, \sigma, \boldsymbol{Y}_p^*$ following an inverse-Gaussian with parameters $(\sqrt{\frac{\lambda_j^2\sigma}{\beta_{pj}^2}}, \lambda_j^2)$, $j = 1, \ldots, k$. The full conditional posterior of $\boldsymbol{\beta}_p$ is obtained by

$$f(\boldsymbol{\beta}_p|\sigma, \boldsymbol{v}, \boldsymbol{\gamma}, \boldsymbol{Y}_p^*, \boldsymbol{X}^*) \propto \exp\{-\frac{1}{2\sigma}[(\boldsymbol{Y}_p^* - \boldsymbol{X}^*\boldsymbol{\beta}_p)^T\boldsymbol{V}^{-1}(\boldsymbol{Y}_p^* - \boldsymbol{X}^*\boldsymbol{\beta}_p) + \boldsymbol{\beta}_p^T\boldsymbol{D}_\gamma^{-1}\boldsymbol{\beta}_p]\}, \tag{3.21}$$

which can be expressed as an $N_k(\bar{\boldsymbol{\mu}}_p, \sigma\bar{\boldsymbol{\Lambda}}^{-1})$, where $\bar{\boldsymbol{\mu}}_p = [\boldsymbol{D}_\gamma^{-1} + \boldsymbol{X}^*\boldsymbol{V}^{-1}\boldsymbol{X}^*]^{-1}\boldsymbol{X}^*\boldsymbol{V}^{-1}\boldsymbol{Y}_p^*$ and $\bar{\boldsymbol{\Lambda}} = \boldsymbol{D}_\gamma^{-1} + \boldsymbol{X}^*\boldsymbol{V}^{-1}\boldsymbol{X}^*$. The full conditional posterior of σ is given by

$$f(\sigma|\boldsymbol{\beta}_p, \boldsymbol{v}, \boldsymbol{\gamma}, \boldsymbol{Y}_p^*, \boldsymbol{X}^*) \propto \sigma^{-(\frac{3n+k+2a_0}{2}+1)} \exp\{-\frac{1}{2\sigma}[(\boldsymbol{Y}_p^* - \boldsymbol{X}^*\boldsymbol{\beta}_p)^T\boldsymbol{V}^{-1}(\boldsymbol{Y}_p^* - \boldsymbol{X}^*\boldsymbol{\beta}_p) \\ + \boldsymbol{\beta}_p^T\boldsymbol{D}_\gamma^{-1}\boldsymbol{\beta}_p + 2p(1-p)\sum_{i=1}^{n} v_i + 2b_0]\}, \tag{3.22}$$

which is an IG distribution with shape $\frac{3n+k+2a_0}{2}$ and scale $\frac{1}{2}[(\boldsymbol{Y}_p^* - \boldsymbol{X}^*\boldsymbol{\beta}_p)^T\boldsymbol{V}^{-1}(\boldsymbol{Y}_p^* - \boldsymbol{X}^*\boldsymbol{\beta}_p) + \boldsymbol{\beta}_p^T\boldsymbol{D}_\gamma^{-1}\boldsymbol{\beta}_p + 2p(1-p)\sum_{i=1}^{n} v_i + 2b_0]$. The full posterior of each

$v_i, i = 1, 2, \ldots, n$ is also tractable:

$$\begin{aligned} f(v_i|\boldsymbol{\beta}_p, \sigma, y_i, \boldsymbol{x}_i) &\propto v_i^{-1/2} \exp\{-\frac{1}{4\sigma}[v_i^{-1}(y_i - (1-2p)v_i - \boldsymbol{x}_i^T\boldsymbol{\beta}_p)^2] - \frac{p(1-p)}{\sigma}v_i\} \\ &= v_i^{-1/2} \exp\{-\frac{1}{4\sigma}[v_i^{-1}(y_i - \boldsymbol{x}_i^T\boldsymbol{\beta}_p)^2 + v_i]\} \\ &= v_i^{-1/2} \exp\{-\frac{1}{2}(v_i^{-1}\bar{\xi}_i^{\,2} + v_i\bar{\zeta}_i^{\,2})\}, \end{aligned} \tag{3.23}$$

where $\bar{\xi}_i^{\,2} = (y_i - \boldsymbol{x}_i^T\boldsymbol{\beta}_p)^2/2\sigma$ and $\bar{\zeta}_i^{\,2} = 1/2\sigma$, which can be recognized as a $GIG(\frac{1}{2}, \bar{\xi}_i, \bar{\zeta}_i)$. A corresponding Gibbs sampling algorithm can be presented as below.

Algorithm 3.4. *Consider a pth $(0 < p < 1)$ Bayesian LASSO regularized QR with prior calibration (3.20) and the big data $\boldsymbol{X}$ and $\boldsymbol{Y}$, we obtain the following Gibbs sampler algorithm.*

Step 1 *the same as presented in* **Algorithm 3.2**.

Step 2 *for each $\boldsymbol{X}_m, \boldsymbol{Y}_m$, a Gibbs sampler for sampling $\boldsymbol{\beta}_p, \sigma$, the $n_m \times 1$ latent vector $\boldsymbol{v}_m$ and $\boldsymbol{\gamma}$ follows the sub-steps below:*

2.1 *denote r as the iteration count. Then set $r = 0$ and establish $(\boldsymbol{\beta}_p^{(r=0)}, \sigma^{(r=0)}, \boldsymbol{v}_m^{(r=0)}, \boldsymbol{\gamma}^{(r=0)})$ to some starting values.*

2.2 *let $\boldsymbol{X}_m^* = \frac{1}{\sqrt{2}}\boldsymbol{X}_m$, $\boldsymbol{Y}_{pm}^* = \frac{1}{\sqrt{2}}(\boldsymbol{Y}_m - (1-2p)\boldsymbol{v}_m), \boldsymbol{V}_m = diag(\boldsymbol{v}_m)$ and $\boldsymbol{D}_\gamma = diag(\boldsymbol{\gamma})$.*

2.3 *follow the inverse-Gaussian conditional posterior of γ_j^{-1}, and the full conditional posteriors of $\boldsymbol{\beta}_p, \sigma, \boldsymbol{v}_m$ given in (3.21) - (3.23),*

(i) sample $\boldsymbol{v}_m^{(r+1)}$ from its GIG posterior $f(\boldsymbol{v}_m|\boldsymbol{\beta}_p^{(r)}, \sigma^{(r)})$.

(ii) sample $\boldsymbol{\gamma}^{(r+1)} = (\gamma_1^{(r+1)}, \ldots, \gamma_k^{(r+1)})^T$, where $1/\gamma_j^{(r+1)}$ follows an inverse-Gaussian with parameters $(\sqrt{\frac{\lambda_j^2\sigma^{(r)}}{(\beta_{pj}^{(r)})^2}}, \lambda_j^2), j = 1, \ldots, k$.

(iii) sample $\sigma^{(r+1)}$ from its IG posterior $f(\sigma|\boldsymbol{\beta}_p^{(r)}, \boldsymbol{v}_m^{(r+1)}, \boldsymbol{\gamma}^{(r+1)})$.

(iv) sample $\boldsymbol{\beta}_p^{(r+1)}$ from its multivariate normal posterior $f(\boldsymbol{\beta}_p|\sigma^{(r+1)}, \boldsymbol{v}_m^{(r+1)}, \boldsymbol{\gamma}^{(r+1)})$.

2.4 *set $r = r + 1$ and return to **Step 2.3** until $r = L$, where L is the number of iteration times.*

Step 3 *calculate the empirical estimates of the means $\bar{\boldsymbol{\beta}}_p, \bar{\sigma}$ and $\bar{\boldsymbol{\gamma}}$ based on the $(L - B)$ realizations of the Gibbs sequence (discarding the first B iterations as a burn-in). Then generate an n_m i.i.d. sample on $\bar{v}_i$, where $\bar{v}_i \sim$*

$GIG(\frac{1}{2}, \bar{\xi}_i, \bar{\zeta}_i)$ with $\bar{\xi}_i^{\,2} = [(y_i - \boldsymbol{x}_i^T \bar{\boldsymbol{\beta}}_p)^2]/2\bar{\sigma}$ and $\bar{\zeta}_i^{\,2} = 1/2\bar{\sigma}, i = 1, 2, \ldots, n_m$. Let $\boldsymbol{D}_\gamma^\dagger = diag(\bar{\boldsymbol{\gamma}}), \boldsymbol{v}_m^\dagger = (\bar{v}_1, \ldots, \bar{v}_{n_m})^T, \boldsymbol{Y}_{pm}^\dagger = \frac{1}{\sqrt{2}}(\boldsymbol{Y}_m - (1-2p)\boldsymbol{v}_m^\dagger)$ and $\boldsymbol{V}_m^\dagger = diag(\boldsymbol{v}_m^\dagger), m = 1, 2, \ldots, M$.

Step 4 *for each subset, the corresponding likelihood can be represented as a form of $NIG_k(\boldsymbol{\mu}_{pm}, \boldsymbol{\Lambda}_m, a_m, b_{pm})$ distribution for $(\boldsymbol{\beta}_p, \sigma)$. Obtain the multiplicative distribution $NIG_k(\boldsymbol{\mu}_p, \boldsymbol{\Lambda}, a, b_p) = \prod_{m=1}^M NIG(\boldsymbol{\mu}_{pm}, \boldsymbol{\Lambda}_m, a_m, b_{pm})$, then the full data posterior is given by merging the prior $NIG_k(\boldsymbol{0}_k, \boldsymbol{D}_\gamma^{-1}, a_0, b_0)$ and the distribution $NIG_k(\boldsymbol{\mu}_p, \boldsymbol{\Lambda}, a, b_p)$:*

$$NIG_k(\bar{\boldsymbol{\mu}}_p, \bar{\boldsymbol{\Lambda}}, \bar{a}, \bar{b}_p) = NIG_k(\boldsymbol{0}_k, \boldsymbol{D}_\gamma^{-1}, a_0, b_0) NIG_k(\boldsymbol{\mu}_p, \boldsymbol{\Lambda}, a, b_p),$$

where $\bar{\boldsymbol{\mu}}_p = [\boldsymbol{D}_\gamma^{-1} + \sum_{m=1}^M \boldsymbol{X}_m^{*T} \boldsymbol{V}_m^{\dagger-1} \boldsymbol{X}_m^*]^{-1} \sum_{m=1}^M \boldsymbol{X}_m^{*T} \boldsymbol{V}_m^{\dagger-1} \boldsymbol{Y}_{pm}^\dagger, \bar{\boldsymbol{\Lambda}} = \boldsymbol{D}_\gamma^{-1} + \sum_{m=1}^M \boldsymbol{X}_m^{*T} \boldsymbol{V}_m^{\dagger-1} \boldsymbol{X}_m^*, \bar{a} = \frac{3n+2a_0}{2}, \bar{b}_p = b_0 + \frac{1}{2}[\sum_{m=1}^M \boldsymbol{Y}_{pm}^{\dagger*T} \boldsymbol{V}_m^{\dagger-1} \boldsymbol{Y}_{pm}^{\dagger*} - \bar{\boldsymbol{\mu}}_p^T \bar{\boldsymbol{\Lambda}} \bar{\boldsymbol{\mu}}_p] + p(1-p) \sum_{m=1}^M \|\boldsymbol{v}_m^\dagger\|_1$ and $\|\cdot\|_1$ denotes the ℓ_1 norm of a vector.

3.7 Numerical demonstrations and real-data analysis

In this section, we assess the performance of the proposed big data based algorithms for posterior distribution calculation through a series of numerical demonstrations and two real-world data analyses.

3.7.1 Numerical demonstrations

3.7.1.1 Bayesian scale mixtures of normals regression

In the Bayesian scale mixtures of normals linear regression scenario, we generate data from a true model of the form $\boldsymbol{Y} = \boldsymbol{X}\boldsymbol{\beta} + \sigma\boldsymbol{\epsilon}$, where $\boldsymbol{Y}$ is a $10^6 \times 1$ response vector, $\boldsymbol{X}$ is a $10^6 \times 10^4$ predictor matrix with the first column assigned as a vector of all 1's and the remaining elements generated from $N(0,1)$. $\boldsymbol{\beta}$ is a $10^4 \times 1$ vector where only the first 10 coefficients $(\beta_0, \ldots, \beta_9)^T = (10, 9, 8, 7, 6, 5, 4, 3, 2, 1)^T$ are set to be non-zero and σ^2 is set as $\sqrt{1.25}$. $\epsilon_i \overset{d}{=} \sqrt{\zeta_i} z_i, i = 1, \ldots, 10^6$ where z_i follows $N(0,1)$ and ζ_i is an independent random variable generated from the uniform distribution $\mathcal{U}(0.5, \sqrt{5})$. We further specify an informative prior $NIG_{10^4}(\boldsymbol{0}, \boldsymbol{I}, 2, 1)$ for $(\boldsymbol{\beta}, \sigma^2)$ where $\boldsymbol{I}$ denotes the identity matrix. The whole data is partitioned into 100 subsets with each containing 10,000 observations. We use R (R Core Team, 2021) as our programming language for all computing tasks. We implement **Algorithm 3.1** and the computation of the specified linear model consumes 6.25 min of CPU time. Table 3.1 reports the posterior means, standard deviations and 95% credible intervals for the non-zero coefficients $(\beta_0, \ldots, \beta_9)^T$. The estimated coefficients for the remaining predictors closely align with the true values of zero, suggesting their decisive exclusion from the regression. The simulation results indicate that our

Parameter	True Value	Mean	Std	95% CI	
				P2.5	P97.5
β_0	10	9.9662	0.0450	9.8769	10.0540
β_1	9	8.9525	0.0466	8.8622	9.0443
β_2	8	8.0115	0.0460	7.9206	8.1023
β_3	7	7.0212	0.0456	6.9320	7.1102
β_4	6	6.0759	0.0435	5.9911	6.1608
β_5	5	4.9944	0.0467	4.9030	5.0864
β_6	4	3.9454	0.0441	3.8582	4.0325
β_7	3	2.9999	0.0463	2.9092	3.0899
β_8	2	1.9993	0.0457	1.9106	2.0897
β_9	1	0.9729	0.0458	0.8829	1.0621

Table 3.1 Estimation results of the first 10 non-zero coefficients for the Bayesian scale mixtures of normals regression model

proposed big data based approach for the Bayesian scale mixtures of normals regression behaves well and provides an accurate estimation of the true regression coefficients.

3.7.1.2 Bayesian quantile regression

To investigate the performance of our proposed algorithms for the pth Bayesian quantile regression, we generate data from a true model $\boldsymbol{Y} = \boldsymbol{X\beta} + \boldsymbol{\epsilon}$, where $\boldsymbol{Y}$ is a $10^6 \times 1$ response vector, $\boldsymbol{X}$ is a $10^6 \times 10^4$ design matrix with all elements generated from $N(0, 1)$. $\boldsymbol{\beta} = (10, 9, 8, \ldots, 1, 0, \ldots, 0)^T$ is a $10^4 \times 1$ vector with only the first 10 coefficients set to be non-zero. $\boldsymbol{\epsilon}$ is the disturbance vector where $\epsilon_i \sim ALD(0, \sigma, p), i = 1, \ldots, 10^6$. We implement **Algorithm 3.2** for our big data BQR model at quantiles $p = 0.50$ and $p = 0.95$ respectively. In each scenario, three postulates of σ are considered: $\sigma = 0.05$, $\sigma = 0.1$ and $\sigma = 0.5$. The given full data is partitioned into 100 subsets with equal size of 10,000 and the Gibbs samplers are run for 15,000 iterations with a burn-in of 5000. An informative g-prior with $g = 100$ is specified, as suggested in Smith and Kohn (1996). The CPU time required for completing parameter estimation in each quantile calibration and sigma postulation is under 2.3 hrs. Tables 3.2-3.4 present the posterior means, standard deviations and 95% credible intervals of the non-zero coefficients for each assignment of σ respectively. Moreover, our approach effectively excludes most predictors with true coefficients equal to zero in the DGP. The acquired estimates validate the efficacy of our proposed big data based algorithms for the BQR model.

3.7.2 Real-data analysis

In this section, we illustrate our divide-and-conquer algorithms for big data Bayesian quantile regression by two real-world data sets.

		p=0.50				p=0.95			
				95% CI				95% CI	
Parameter	True Value	Mean	Std	P2.5	P97.5	Mean	Std	P2.5	P97.5
β_0	10	9.8967	0.0407	9.8156	9.9756	9.7613	0.0503	9.6625	9.8574
β_1	9	8.8896	0.0375	8.8158	8.9639	9.0528	0.0485	8.9580	9.1486
β_2	8	7.9260	0.0402	7.8469	8.0043	7.8225	0.0483	7.7269	7.9165
β_3	7	6.9322	0.0365	6.8594	7.0033	6.7894	0.0457	6.6983	6.8789
β_4	6	5.9569	0.0363	5.8866	6.0277	5.7217	0.0439	5.6365	5.8077
β_5	5	4.9479	0.0342	4.8809	5.0162	5.0657	0.0451	4.9781	5.1542
β_6	4	3.9637	0.0331	3.8996	4.0281	3.9624	0.0373	3.8909	4.0356
β_7	3	2.9587	0.0352	2.8903	3.0291	2.9109	0.0417	2.8298	2.9932
β_8	2	1.9797	0.0341	1.9130	2.0443	1.9608	0.0410	1.8793	2.0406
β_9	1	0.9905	0.0307	0.9293	1.0512	1.0409	0.0410	0.9615	1.1209

Table 3.2 Estimation results of the first 10 non-zero coefficients for the Bayesian quantile regression model for $\sigma = 0.05$

		p=0.50				p=0.95			
				95% CI				95% CI	
Parameter	True Value	Mean	Std	P2.5	P97.5	Mean	Std	P2.5	P97.5
β_0	10	9.9201	0.0429	9.8355	10.0025	9.6914	0.1711	9.3500	10.0240
β_1	9	8.9235	0.0396	8.8461	9.0017	8.8140	0.1718	8.4757	9.1524
β_2	8	7.9363	0.0391	7.8596	8.0129	8.0678	0.1708	7.7326	8.4005
β_3	7	6.9333	0.0376	6.8590	7.0063	6.9045	0.1611	6.5836	7.2187
β_4	6	5.9282	0.0331	5.8638	5.9925	5.6809	0.1565	5.3736	5.9869
β_5	5	4.9604	0.0386	4.8859	5.0361	4.8938	0.1718	4.5582	5.2296
β_6	4	3.9523	0.0339	3.8860	4.0183	3.7937	0.1547	3.4907	4.0929
β_7	3	2.9767	0.0354	2.9072	3.0461	3.0477	0.1616	2.7333	3.3663
β_8	2	1.9761	0.0341	1.9092	2.0426	2.0570	0.1624	1.7381	2.3724
β_9	1	0.9944	0.0322	0.9311	1.0570	1.0894	0.1606	0.7765	1.4029

Table 3.3 Estimation results of the first 10 non-zero coefficients for the Bayesian quantile regression model for $\sigma = 0.1$

3.7.2.1 Airline on-time performance data

The airline on-time performance data set from the 2009 ASA Data Expo is publicly available at http://stat-computing.org/dataexpo/2009/the-data.html. The data has been used for a demonstration of massive data by Wang et al. (2016a) and Schifano et al. (2016). It consists of flight arrival and departure details for all commercial flights within the United States from October 1987 to April 2008. About 12 million flights were involved with 29 variables. Due to the computing limit, we only consider a complete sub-dataset of the year 2008 with $N = 584,583$ after removing all the missing records. We consider arrival

		p=0.50				p=0.95			
				95% CI				95% CI	
Parameter	True Value	Mean	Std	P2.5	P97.5	Mean	Std	P2.5	P97.5
β_0	10	9.8176	0.0549	9.7084	9.9242	8.6045	0.2124	8.1819	9.0229
β_1	9	8.8578	0.0509	8.7589	8.9583	7.6146	0.2078	7.2084	8.0211
β_2	8	7.9462	0.0514	7.8440	8.0468	6.8295	0.2068	6.4228	7.2320
β_3	7	6.9935	0.0498	6.8934	7.0900	6.1102	0.2028	5.7050	6.5057
β_4	6	6.0137	0.0467	5.9227	6.1041	5.4773	0.1937	5.0990	5.8537
β_5	5	5.0526	0.0488	4.9577	5.1492	4.6158	0.2114	4.2051	5.0316
β_6	4	3.9439	0.0389	3.8693	4.0206	3.9798	0.1962	3.5949	4.3631
β_7	3	2.9775	0.0458	2.8884	3.0677	2.0863	0.2041	1.6877	2.4921
β_8	2	2.1202	0.0451	2.0300	2.2076	1.4316	0.2028	1.0351	1.8261
β_9	1	1.0292	0.0433	0.9451	1.1135	0.8008	0.2015	0.4081	1.1914

Table 3.4 Estimation results of the first 10 non-zero coefficients for the Bayesian quantile regression model for $\sigma = 0.5$

delay (AD) as a continuous variable by modelling $\log(AD - \min(AD) + 1)$ and employ a linear model that specifies the pth quantile of AD as follows:

$$Q_p(AD) = \beta_{p0} + \beta_{p1}HD + \beta_{p2}DIS + \beta_{p3}NF + \beta_{p4}WF + \epsilon,$$

where HD is the departure time (continuous, in hours), DIS is the distance (continuous, in thousands of miles), NF is the day/night flight indicator (binary; 1 if departure between 8 p.m. and 5 a.m., 0 otherwise) and WF is the weekend/weekday flight indicator (binary; 1 if departure occurred during the weekend, 0 otherwise). This model was also investigated by Schifano et al. (2016).

We fit our big data BQR to the above specified regression model by implementing **Algorithm 3.2** at $p = 0.50$, 0.75 and 0.95 respectively. In each scenario, the whole observations are partitioned into 100 subsets with the size of $n_m = 5845$ for $m = 1, \ldots, 99$ and $n_{100} = 5928$. We assign the informative g-prior by choosing $g = 100$. All results are based on 15,000 draws obtained from the Gibbs samplers with a burn-in of 5000 iterations. Table 3.5 presents the estimated coefficients and posterior standard deviations at the specified quantile levels. We observe that the departure time bears a positive association with the arrival delay, whereas the distance, night-time and weekend flights have negative effects on the delay across all the three quantiles considered. Nevertheless, the effects of these covariates are mitigated with the increase of quantile. Night-time flight is found to be a non-negligible factor to improve on-time performance of flights facing median and long arrival delays. This empirical study shows that our proposed BQR method facilitates the investigation of the effects of different factors on various levels of flight arrival delays in the big data scenario.

	$p = 0.50$		$p = 0.75$		$p = 0.95$	
	Coeff	Std	Coeff	Std	Coeff	Std
Intercept	1.9483	3.3380	2.6819	3.3598	4.1028	2.9804
HD	0.0790	0.2038	0.0735	0.2014	0.0403	0.1709
DIS	-0.0577	1.5080	-0.0573	1.5440	-0.0150	1.4152
NF	-0.4222	3.0845	-0.3932	3.0592	-0.1398	2.6500
WF	-0.0545	1.9676	-0.0444	1.9923	-0.0372	1.8048

Table 3.5 Coefficient estimates and posterior standard deviations ($\times 10^3$) of big data BQR estimator for the airline on-time data

	Aeolos	Iweco	Rokas
Min	0.000	0.000	0.000
Quantile (.25)	1.692	0.921	1.573
Median	4.002	2.112	4.579
Mean	4.142	2.141	4.857
Quantile (.75)	6.745	3.426	8.049
Max	8.302	4.549	11.635
Standard deviation	2.649	1.346	3.407
Sample size	17819	15621	21949

Table 3.6 Summary statistics for wind power observations at Aeolos, Iweco and Rokas

3.7.3 Hourly wind power data

The hourly wind power data set is recorded from 31 December 2007 to 30 December 2010 at the following three wind farms in Crete: Aeolos, Iweco and Rokas. The data is a collection of hourly observations for wind speed (measured in m/s), direction (measured in degrees) and power (measured in megawatts). A complete wind power data of the year 2010 is examined in Taylor (2017). We remove all the missing data and retain positive observations of the recorded hourly periods. Table 3.6 presents the summary statistics for wind power observations (in MW) at Aeolos, Iweco and Rokas respectively.

We fit our big data BQR by modeling the wind power as a linear function of wind speed and direction. We implement **Algorithm 3.2** for these three power sequences at $p = 0.50$ and $p = 0.95$ respectively. In each case, the Gibbs samplers are run for 11000 iterations, discarding the first 1000 as a burn-in. For Aeolos farm, the whole observations are partitioned into 50 subsets with the size of $n_1 = n_2 \ldots = n_{49} = 356$ and $n_{50} = 375$. For Iweco, we partition the whole data into 50 subsets with the size of $n_1 = n_2 \ldots = n_{49} = 312$ and $n_{50} = 333$. For Rokas, we consider 50 subsets as $n_1 = n_2 \ldots = n_{49} = 438$ and $n_{50} = 487$. We assign the informative g-prior by choosing $g = 100$. Table 3.7 displays the

	Aeolos				Iweco				Rokas			
Model Covariates	$p = 0.50$		$p = 0.95$		$p = 0.50$		$p = 0.95$		$p = 0.50$		$p = 0.95$	
	Coeff	S.D.	Coeff	S.D.	Coeff	S.D.	Coeff	S.D.	Coeff	S.D.	Coeff	S.D.
Intercept	-2.8624	0.0151	-3.6681	0.0160	-0.7907	0.0110	-0.5663	0.0135	-2.8004	0.0130	-1.9270	0.0150
Speed	0.7485	0.0151	1.0447	0.0018	0.3770	0.0015	0.5316	0.0015	0.7860	0.0013	1.0616	0.0010
Direction	-0.0003	0.0000	-0.0021	0.0000	-0.0023	0.0000	-0.0039	0.0000	0.0005	0.0000	-0.0040	0.0000

Table 3.7 Coefficient estimates along with posterior standard deviations (S.D.) for Aeolos, Iweco and Rokas in big data BQR analysis

estimates and posterior standard deviations in our big data BQR model for the given three wind power series separately. Note that for all power series, the estimated coefficients of direction are close to zero at the measured percentiles, meaning that the effect of wind direction on power seems to be minor. Instead, wind power presents a much stronger correlation to speed than to direction. The positive coefficients of speed indicate that as wind speed increases, so does the power capacity. Furthermore, it is visible that speed has a greater impact on higher (95th percentile) power than lower (50th percentile) power capacity for all the three aforementioned wind farms.

3.8 Chapter summary

The methods of Bayesian scale mixtures of normals linear regression and Bayesian quantile regression for big data analysis, including variable selection and posterior predictive distributions, have been explored. This is achieved by using ALD-based working likelihood functions and conjugate NIG priors. The resulting algorithms are easily implemented and the numerical demonstrations show that the proposed approaches are promising.

Chapter 4

Bayesian two-part Kumaraswamy quantile mixed model with latent class for semi-continuous longitudinal data

4.1 Introduction

Semi-continuous data frequently exhibit a combination of zero values and positively distributed values. These variables can be conceptualized as outcomes of two distinct stochastic processes. One process governs the occurrence of zeros, while the other determines the actual values for non-zero observations. This conceptual framework gives rise to a two-part model designed to account for both the prevalence of zeros and the typically skewed distribution of non-zero observations (Heckman, 1976).

The literature focusing on two-part models targeted at semi-continuous data, where the continuous component is characterized by variables with bounded support, has garnered increasing attention. In instances where the values of variables fall within the standard unit interval, the predominant model employed is the beta regression model proposed by Ferrari and Cribari-Neto (2004). Despite the beta distribution being the premier family of continuous distributions on bounded support and demonstrating reasonable tractability, it possesses certain complexities. Specifically, its distribution function is an incomplete beta function ratio and its quantile function is the inverse thereof.

Jones (2009) introduced an alternative two-parameter distribution defined on the interval $(0, 1)$, termed as the Kumaraswamy distribution which is characterized by two positive shape parameters. The roots of this distribution can be traced back to hydrological studies (Kumaraswamy, 1980; Nadarajah, 2008). However, its close interchangeability with the beta distribution has not garnered widespread recognition until Jones (2009), where a comprehensive analysis of the properties of the Kumaraswamy distribution is first provided to the existing literature. Notably, the density of the Kumaraswamy distribution exhibits similar characteristics to the beta distribution, such as unimodality, uniantimodality, monotonicity and constancy, depending on the parameter values. Both the Kumaraswamy and beta distributions exhibit favourable behaviour of skewness and kurtosis measures as functions of the respective parameters, showcasing their shared attributes. However, Jones (2009) proposed the Kumaraswamy distribution as a viable alternative to the beta distribution, emphasizing that the former shares many properties with the latter while offering greater ease of handling in various aspects. A distinctive feature of the Kumaraswamy distribution is its straightforward explicit formula for the distribution function and quantile functions, eliminating dependencies on special functions. This simplicity facilitates effortless random variate generation and plays a specific role in quantile-based statistical modelling, making the Kumaraswamy distribution particularly appealing for practical applications due to its tractability. Additionally, the distribution provides simpler formulae for moments of order statistics compared to the beta distribution. Bayes et al. (2017) proposed a new quantile parametric mixed regression model which is built upon a reparameterization of the Kumaraswamy distribution in terms of a given quantile and the precision parameter, and formulated a Bayesian approach for parameter inference including model comparison criteria. The investigation of the beta and Kumaraswamy regression models, while accounting for the challenge of multi-collinearity, is explored in Pirmohammadi and Bidram (2022). Performance evaluations of these models are conducted via the maximum likelihood estimator and the Liu Regression Estimator (Liu, 1993). Through a comprehensive comparative analysis, the authors demonstrated that the Kumaraswamy regression model outperforms the beta model in both simulation scenarios and real-world data analyses involving the gasoline yield data and data from the Australian Institute of Sport.

The Kumaraswamy and beta distributions are unsuitable for datasets containing zeros and/or ones. In light of these limitations, various adaptations in the realm of semi-continuous data modelling with inflated beta and Kumaraswamy distributions have emerged. For example, Cook et al. (2008) developed a zero-inflated beta model with the analysis of corporate capital structure decisions. Ospina and Ferrari (2010) introduced the zero-and-one-inflated beta distribution and discussed its estimation based on maximum likelihood and conditional moments. Liu and Eugenio (2018) presented a thorough review and comparison of Bayesian and likelihood-based inferences in beta regression and zero-or-one-inflated beta regression. More recently, Cribari-Neto and Santos (2019) introduced inflated Kumaraswamy distributions. Bayer et al. (2021) proposed an inflated version of the Kumaraswamy distribution utilizing the median-based parametrization from Mitnik and Baek (2013) and conducted model parameter estimation by maximum likelihood inference.

In the context of longitudinal data analysis, the random effects can account for heterogeneity that is inexplicable by fixed effects estimator. However, they may become inappropriate when the response profile exhibits underlying subpopulations with different patterns. Latent class models (LCMs) are becoming increasingly popular to handle such heterogeneous processes. A major analytic goal of LCMs is to identify subgroups, called latent classes, based upon responses to a set of observed covariates. LCMs assume that subjects are assigned into a finite number of classes which share similar characteristics of data. As such, individuals have their own unique longitudinal trajectories signified by a set of random effects with variance parameters varying across distinct clusters. The latent class approach has become a widely applicable instrument for detecting and decomposing unobserved heterogeneity in longitudinal data analysis. Lin et al. (2002) employed the latent class joint model to the nutritional prevention of cancer (NPC) trials data to aid the estimation of longitudinally measured prostate-specific antigen (PSA) trajectories and risk for prostate cancer in distinct subpopulations. Han et al. (2007) proposed a parametric latent class model to jointly accommodate association between the biomarker and recurrent event processes. Proust-Lima et al. (2014) developed joint latent class models for predicting prostate cancer recurrence after radiation therapy based on repeated measures of Prostate Specific Antigen. In the Bayesian framework, Elliott et al. (2005) considered a Bayesian latent growth curve model to identify patient trajectories of positive affect and negative events over a 35-day experiment in psychiatric care settings. Leiby et al. (2009) proposed a Bayesian growth curve latent class factor analytic model for multivariate data from a randomized clinical trial evaluating the efficacy of a new treatment for interstitial cystitis. Neelon et al. (2011a) put forth a Bayesian two-part latent class model to delineate the impact of parity on mental health utilization and expenditures. In the same year, Neelon et al. (2011b) pioneered a flexible Bayesian growth mixture model to jointly investigate the associations between longitudinal blood pressure measurements, preterm birth (PTB) and low birth weight (LBW). Very recently, Yang and Puggioni (2021) explored Bayesian latent class models for longitudinal data with zero-inflated count response variables and applied the established method to cigarette smoking data. Kim et al. (2023) developed a Bayesian latent class modelling approach for characterizing variation in circadian rhythms among longitudinal metabolites data. To our knowledge, there is a sparse literature on the Bayesian analysis of two-part latent class quantile models for bounded longitudinal data.

This chapter extends the model introduced by Bayes et al. (2017) to a two-part latent class Kumaraswamy quantile mixed regression with Bayesian inference for semi-continuous longitudinal data. The binomial component is specified via mixed effects probit regression and the continuous component is formulated through a Kumaraswamy quantile mixed effects model. The two components are linked together by correlated random effects since ignoring such potential correlation can yield biased inferences (Su et al., 2009). We employ the proposed model to investigate how available covariates affect the proportion of outpatient expenses to the total spending on health services in US families. The well-known RAND Health Insurance Experiment serves as the source of information for the real data analysis presented in Section 4.4. In view of the tendency among enrollees to share characteristics related to medical care spending, the two-part latent class quantile mixed model permits us to cluster individuals with similar observed expenditure trajectories and identify latent classes of subjects who exhibited moderate-to-high probability of outpatient spending and the amount spent as well as subjects who were hesitant to such expenses across the total health outlay.

4.2 Bayesian two-part latent class Kumaraswamy quantile mixed model

4.2.1 Two-part latent class Kumaraswamy quantile mixed model

A random variable Y follows the Kumaraswamy distribution if its probability density function is given by

$$f(y|\alpha,\beta) = \alpha\beta y^{\alpha-1}(1-y^{\alpha})^{\beta-1},\ 0<y<1,\ \alpha,\beta>0. \tag{4.1}$$

The cumulative distribution function (cdf) has a closed expression by

$$F(y|\alpha,\beta) = 1-(1-y^{\alpha})^{\beta} \tag{4.2}$$

Then the quantile function of the Kumaraswamy distribution is readily obtained from the cdf as

$$\kappa(q) = F^{-1}(q) = \{1-(1-q)^{1/\beta}\}^{1/\alpha} \tag{4.3}$$

for any quantile level $0<q<1$. Following Bayes et al. (2017), we consider a reparameterization of the Kumaraswamy distribution in terms of the q-th quantile $\kappa = \kappa(q)$ and the precision parameter $\varphi = \varphi(q)$

$$\kappa = \{1-(1-q)^{1/\beta}\}^{1/\alpha} \text{ and } \varphi = -\log(1-(1-q)^{1/\beta}), \tag{4.4}$$

where q is assumed to be known and the parameter space of $(\kappa,\varphi)^T$ is given by $(0,1)\times(0,\infty)$. Then the pdf and the cdf of the reparameterized Kumaraswamy distribution turn out to be

$$f(y|\kappa,\varphi) = -\frac{\log(1-q)\varphi}{\log(1-e^{-\varphi})\log(\kappa)} y^{-\frac{\varphi}{\log(\kappa)}-1}\{1-y^{-\frac{\varphi}{\log(\kappa)}}\}^{\frac{\log(1-q)}{\log(1-e^{-\varphi})}-1} \tag{4.5}$$

and

$$F(y|\kappa, \varphi) = 1 - \{1 - y^{-\frac{\varphi}{\log(\kappa)}}\}^{\frac{\log(1-q)}{\log(1-e^{-\varphi})}} \tag{4.6}$$

respectively. We denote this parameterization as $Y \sim \text{Kum}(\kappa, \varphi, q)$ where the quantile parameter $\kappa \in (0,1)$ acts as a location parameter, $\varphi > 0$ is a precision parameter and the probability q is assumed to be fixed according to the quantile of interest.

Let y_{ij} denote the semi-continuous response of the j-th repeated measurement for the i-th individual, where $y_{ij} \in [0,1), i = 1, \ldots, n$ and $j = 1, \ldots, n_i$. The density of y_{ij} in the proposed two-part Kumaraswamy quantile mixed model can be written in the following way:

$$f(y_{ij}) = (1 - \gamma_{ij})\mathcal{I}(y_{ij} = 0) + \gamma_{ij} \times \text{Kum}(y_{ij}|\kappa_{ij}, \varphi, q)\mathcal{I}(0 < y_{ij} < 1), \tag{4.7}$$

where $\gamma_{ij} = P_r(y_{ij} > 0)$, $\mathcal{I}(\cdot)$ is the indicator function and $\text{Kum}(\kappa_{ij}, \varphi, q)$ denotes the Kumaraswamy distribution with observation-specific quantile κ_{ij}, precision φ and fixed quantile level q. The m-th raw moment of this distribution is given by $\text{E}(Y_{ij}^m|\kappa_{ij}, \varphi, q, \gamma_{ij}) = \gamma_{ij}\mu_m$, where μ_m is the m-th raw moment of the distribution $\text{Kum}(\kappa_{ij}, \varphi, q)$. For example, we have

$$\begin{aligned} \text{E}(Y_{ij}|\kappa_{ij}, \varphi, q, \gamma_{ij}) &= \gamma_{ij}\mu_1, \\ \text{Var}(Y_{ij}|\kappa_{ij}, \varphi, q, \gamma_{ij}) &= \gamma_{ij}\mu_2 - \gamma_{ij}^2\mu_1^2. \end{aligned}$$

The model (4.7) can be extended to accommodate latent classes by introducing a latent categorical variable C_i such that $C_i = k$ if subject i belongs to class k, $k = 1, \ldots, K$. Given the value of C_i, the density of Y_{ij} is obtained with class-specific parameters:

$$\begin{aligned} f(y_{ij}|C_i = k) &= (1 - \gamma_{ijk})\mathcal{I}(y_{ij} = 0) + \gamma_{ijk} \times \text{Kum}(y_{ij}|\kappa_{ijk}, \varphi_k, q)\mathcal{I}(0 < y_{ij} < 1), \\ g_\gamma(\gamma_{ijk}) &= \boldsymbol{x}_{1ij}^T\boldsymbol{\beta}_{1k} + b_{1i}, \\ g_\kappa(\kappa_{ijk}) &= \boldsymbol{x}_{2ij}^T\boldsymbol{\beta}_{2k} + b_{2i}, \;\; k = 1, \ldots, K. \end{aligned} \tag{4.8}$$

We refer to the formulation (4.8) as a two-part latent class Kumaraswamy quantile mixed model, where g_γ and g_κ denote the probit or logit link function for γ_{ijk} and κ_{ijk} respectively; $\boldsymbol{x}_{lij}$ are $p_l \times 1$ vectors of fixed effect covariates for component l; $\boldsymbol{\beta}_{lk}$ are class-specific fixed effect coefficient vectors for component l $(l = 1, 2)$ and $\boldsymbol{b}_i|C_i = (b_{1i}, b_{2i})^T|C_i$ is a stacked random effects vector for subject i, here we assume $\boldsymbol{b}_i|C_i \sim N_2(\boldsymbol{0}, \boldsymbol{\Sigma}_k)$ with class-specific covariance $\boldsymbol{\Sigma}_k$.

We assume the latent categorical variable C_i follows a categorical distribution, that is, C_i takes the value k with probability δ_{ik}, $k = 1, \ldots, K$:

$$\begin{aligned} C_i &\sim \text{Cat}(\delta_{1k}, \ldots, \delta_{iK}), \\ \delta_{ik} &= \frac{e^{\boldsymbol{\eta}_i^T\boldsymbol{\phi}_k}}{\sum_{h=1}^K e^{\boldsymbol{\eta}_i^T\boldsymbol{\phi}_h}}, \end{aligned} \tag{4.9}$$

where δ_{ik} is formulated via a generalized logit model, $\boldsymbol{\eta}_i$ denotes a $v \times 1$ covariate vector and $\boldsymbol{\phi}_k$ represents the vector of associated regression parameters for each class k, with $\boldsymbol{\phi}_1 = \mathbf{0}$ for identifiability such that the remaining coefficients can be interpreted in terms of the change in log-odds relative to this reference category. We assume the number of latent classes K is fixed throughout this chapter. For model comparison and the determination of K, we consider the deviance information criterion (DIC) proposed by Spiegelhalter et al. (2002) and its adaption DIC$_3$ developed by Celeux et al. (2006).

4.2.2 Bayesian inference for the model

4.2.2.1 Prior specification

We assign weakly informative priors to all class-specific parameters $\{\boldsymbol{\phi}_k, \boldsymbol{\beta}_{1k}, \boldsymbol{\beta}_{2k}, \varphi_k, \boldsymbol{\Sigma}_k\}$. Specifically, we impose a diffuse v-dimensional normal prior $\pi(\boldsymbol{\phi}_k) = N_v(\mathbf{0}, 100\boldsymbol{I}_v)$ on $\boldsymbol{\phi}_k$. In the case of $v = 1$, we assign a conjugate Dirichlet prior $\text{Dir}(K, \boldsymbol{\alpha})$ on probabilities $\boldsymbol{\delta} = (\delta_1, \dots, \delta_K)^T$ directly, where $\boldsymbol{\alpha} = (\alpha_1, \dots, \alpha_K)^T$ is a collection of the concentration hyperparameters for the Dirichlet distribution. Then the conditional posterior of $\boldsymbol{\delta}$ can be given by $\boldsymbol{\delta}|C_i, \boldsymbol{\alpha} \sim \text{Dir}(K, \boldsymbol{m} + \boldsymbol{\alpha})$, where $\boldsymbol{m} = (m_1, \dots, m_K)^T$ with m_k denoting the number of occurrences of $C_i = k$, $k = 1, \dots, K$. The hyperparameter α_k plays a crucial role in influencing how uniform the distribution will be. Specifically, it denotes the proportion of cases allocated to each of the K latent classes. When the values of α_k are equal across all classes, the estimated class proportions are more evenly balanced in size. The choice of α_k can significantly impact the posterior results. Research has indicated that an increase in α_k leads to more informative Dirichlet distributions, thereby reducing the likelihood of very small mixture weights. This reduction in probability mass assigned by the prior to membership vectors with empty components contributes to a lower probability of overestimating the number of latent classes K in the dataset (Frühwirth-Schnatter, 2006; Nobile, 2004). Adhering to the suggestion in Asparouhov and Muthén (2011), we adopt the practice of using values of α_k larger than 1 (e.g., $\alpha_k = 10$) to prevent the emergence of empty class solutions.

For the fixed effect coefficients, we assume priors $\pi(\boldsymbol{\beta}_{1k}) = N_{p_1}(\boldsymbol{\mu}_{\beta_1}, \boldsymbol{\Sigma}_{\beta_1})$ and $\pi(\boldsymbol{\beta}_{2k}) = N_{p_2}(\boldsymbol{\mu}_{\beta_2}, \boldsymbol{\Sigma}_{\beta_2})$ respectively. Above, we have assumed identical prior hyperparameters throughout different classes, but in general this is not necessary. We assign noninformative inverse gamma priors $\mathcal{IG}(\rho_1, \rho_2)$ with shape ρ_1 and scale ρ_2 for the Kumaraswamy precisions φ_k. For class-specific covariance $\boldsymbol{\Sigma}_k$ of the random effects vector $\boldsymbol{b}_i$, we assume a conjugate inverse-Wishart prior $\mathcal{IW}(\tau_0, \boldsymbol{\Psi}_0)$ with degrees of freedom τ_0 and 2×2 scale matrix $\boldsymbol{\Psi}_0$, which leads to closed-form conditional inverse-Wishart posteriors.

4.2.2.2 Posterior formulation

Denote $\boldsymbol{\theta}_k = (\boldsymbol{\beta}_{1k}^T, \boldsymbol{\beta}_{2k}^T, \varphi_k)^T$ and assume a priori independence of the parameters, the joint posterior of the full model parameters is obtained by

$$
\begin{aligned}
& f(\boldsymbol{\phi}_1^T, \ldots, \boldsymbol{\phi}_K^T, C_1, \ldots, C_n, \boldsymbol{\theta}_1^T, \ldots, \boldsymbol{\theta}_K^T, \boldsymbol{b}_1^T, \ldots, \boldsymbol{b}_n^T, \boldsymbol{\Sigma}_1, \ldots, \boldsymbol{\Sigma}_K | \boldsymbol{y}) \\
& \propto \prod_{k=1}^{K} \left\{ \prod_{i=1}^{n} \Big[\prod_{j=1}^{n_i} f(y_{ij}|\boldsymbol{\theta}_k, \boldsymbol{b}_i)\, \delta_{ik}\, f(\boldsymbol{b}_i|\boldsymbol{\Sigma}_k) \Big]^{\mathcal{I}(C_i=k)} \times \pi(\boldsymbol{\beta}_{1k})\, \pi(\boldsymbol{\beta}_{2k})\, \pi(\boldsymbol{\Sigma}_k) \right\} \prod_{h=2}^{K} \pi(\boldsymbol{\phi}_h),
\end{aligned}
\tag{4.10}
$$

where $\boldsymbol{y} = (y_{11}, \ldots, y_{1n_1}, \ldots, y_{n1}, \ldots, y_{nn_n})^T$ and $f(y_{ij}|\boldsymbol{\theta}_k, \boldsymbol{b}_i)$ is given in (4.8). We elaborate the posterior formulations of $\boldsymbol{\phi}_k, C_i, \boldsymbol{\theta}_k, \boldsymbol{b}_i$ and $\boldsymbol{\Sigma}_k$ as follows:

(i) The full conditional for the v-dimensional vector $\boldsymbol{\phi}_k, k = 1, \ldots, K$ is given by

$$
\begin{aligned}
f(\boldsymbol{\phi}_k|\cdot) & \propto \prod_{i=1}^{n} P_r(C_i = k|\boldsymbol{\phi}_k)^{I(C_i=k)} \pi(\boldsymbol{\phi}_k) \\
& = \prod_{i:C_i=k} \left(\frac{e^{\boldsymbol{\eta}_i^T \boldsymbol{\phi}_k}}{\sum_{h=1}^{K} e^{\boldsymbol{\eta}_i^T \boldsymbol{\phi}_h}} \right) N_v(\mathbf{0}, 100\boldsymbol{I}_v).
\end{aligned}
\tag{4.11}
$$

Considering there is no closed-form expression of the density (4.11), we resort to a random-walk Metropolis approach for sampling this full conditional posterior, where a multivariate v-dimensional normal proposal distribution centered at previous value $\boldsymbol{\phi}_k^{\text{old}}$ and with covariance matrix $s_v\, \boldsymbol{\Sigma}_{\phi_k}$ is employed. For better mixing and correct convergence, we adopt the idea of adaptive Metropolis algorithm proposed by Haario et al. (1999, 2001) to tune $\boldsymbol{\Sigma}_{\phi_k}$ using the empirical covariance obtained from an extended burn-in period. The use of the scaling parameter s_v is to achieve an optimal acceptance rate of approximately 23%. As a basic choice, we adopt the value $s_v = 2.4/\sqrt{v}$ as suggested in Gelman et al. (1996).

(ii) We draw the latent classification variable C_i from its full categorical posterior distribution with updated probabilities $\bar{\delta}_{ik}$:

$$
\begin{aligned}
C_i & \sim \text{Cat}(\bar{\delta}_{1k}, \ldots, \bar{\delta}_{iK}), \\
\bar{\delta}_{ik} & = \frac{\delta_{ik} \big[\prod_{j=1}^{n_i} f(y_{ij}|\boldsymbol{\theta}_k, \boldsymbol{b}_i) \big] \pi(\boldsymbol{b}_i|C_i = k)}{\sum_{h=1}^{K} \delta_{ih} \big[\prod_{j=1}^{n_i} f(y_{ij}|\boldsymbol{\theta}_h, \boldsymbol{b}_i) \big] \pi(\boldsymbol{b}_i|C_i = h)} \\
& = \frac{\delta_{ik} \big[\prod_{j=1}^{n_i} f(y_{ij}|\boldsymbol{\theta}_k, \boldsymbol{b}_i) \big] N_2(\mathbf{0}, \boldsymbol{\Sigma}_k)}{\sum_{h=1}^{K} \delta_{ih} \big[\prod_{j=1}^{n_i} f(y_{ij}|\boldsymbol{\theta}_h, \boldsymbol{b}_i) \big] N_2(\mathbf{0}, \boldsymbol{\Sigma}_h)}.
\end{aligned}
\tag{4.12}
$$

(iii) For the fixed effect coefficient vector $\boldsymbol{\beta}_{1k}$ of the component 1 in model (4.8), we consider a probit link g_γ for modelling γ_{ijk}, that is, $P_r(y_{ij} > 0|C_i = k, \boldsymbol{\beta}_{1k}, b_{1i}) = \gamma_{ijk} = \Phi(\boldsymbol{x}_{1ij}^T \boldsymbol{\beta}_{1k} + b_{1i})$, where $\Phi(\cdot)$ denotes the cdf of a standard normal distribution. Assuming a multivariate normal prior $\pi(\boldsymbol{\beta}_{1k}) = N_{p1}(\boldsymbol{\mu}_{\beta_1}, \boldsymbol{\Sigma}_{\beta_1})$, the posterior of $\boldsymbol{\beta}_{1k}$ is given by

$$f(\boldsymbol{\beta}_{1k}|\cdot) \propto \Big\{ I(y_{ij} > 0) f(y_{ij}|C_i = k, \boldsymbol{\beta}_{1k}, b_{1i})\, \Phi(\boldsymbol{x}_{1ij}^T \boldsymbol{\beta}_{1k} + b_{1i}) \\ + I(y_{ij} = 0)[1 - \Phi(\boldsymbol{x}_{1ij}^T \boldsymbol{\beta}_{1k} + b_{1i})] \Big\}\, \pi(\boldsymbol{\beta}_{1k}).$$

Due to the intractability of this posterior density, we resort to the well-known data augmentation strategy of the Bayesian probit regression model proposed by Albert and Chib (1993), which enables Gibbs-sampling for the model fitting. Specifically, we introduce latent variables $z_{ij}, i = 1, \ldots, n$, $j = 1, \ldots, n_i$ such that $z_{ij}|C_i = k, \boldsymbol{\beta}_{1k}, b_{1i} \sim N(\boldsymbol{x}_{1ij}^T \boldsymbol{\beta}_{1k} + b_{1i}, 1)$. The augmented likelihood is thus given by

$$f(y_{ij}, z_{ij}|C_i = k, \boldsymbol{\beta}_{1k}, b_{1i}) = \prod_{i:C_i=k} \prod_{j=1}^{n_i} N(z_{ij}|\boldsymbol{x}_{1ij}^T \boldsymbol{\beta}_{1k} + b_{1i}, 1) \times \\ \big[I(z_{ij} > 0) I(y_{ij} > 0) f(y_{ij}|C_i = k, \boldsymbol{\beta}_{1k}, b_{1i}) + I(z_{ij} \leqslant 0) I(y_{ij} = 0) \big].$$

The full conditional distribution of $z_{ij}|y_{ij}, C_i = k, \boldsymbol{\beta}_{1k}, b_{1i}$ follows a truncated normal distribution:

$$f(z_{ij}|y_{ij} > 0, C_i = k, \boldsymbol{\beta}_{1k}, b_{1i}) \propto N(z_{ij}|\boldsymbol{x}_{1ij}^T \boldsymbol{\beta}_{1k} + b_{1i}, 1)\, I(z_{ij} > 0), \\ f(z_{ij}|y_{ij} = 0, C_i = k, \boldsymbol{\beta}_{1k}, b_{1i}) \propto N(z_{ij}|\boldsymbol{x}_{1ij}^T \boldsymbol{\beta}_{1k} + b_{1i}, 1)\, I(z_{ij} \leqslant 0).$$

The full conditional of $\boldsymbol{\beta}_{1k}$ follows a p_1-dimensional normal distribution:

$$\begin{aligned} f(\boldsymbol{\beta}_{1k}|\cdot) &= N_{p_1}(\bar{\boldsymbol{\mu}}_{\beta_{1k}}, \bar{\boldsymbol{\Sigma}}_{\beta_{1k}}), \\ \text{where} \quad \bar{\boldsymbol{\Sigma}}_{\beta_{1k}} &= (\boldsymbol{X}_k^T \boldsymbol{X}_k + \boldsymbol{\Sigma}_{\beta_1}^{-1})^{-1}, \\ \bar{\boldsymbol{\mu}}_{\beta_{1k}} &= \bar{\boldsymbol{\Sigma}}_{\beta_{1k}} \big[\boldsymbol{X}_k^T (\boldsymbol{z}_k - \boldsymbol{b}_{1k}) + \boldsymbol{\Sigma}_{\beta_1}^{-1} \boldsymbol{\mu}_{\beta_1} \big]. \end{aligned} \tag{4.13}$$

Here $\boldsymbol{\mu}_{\beta_1}$ and $\boldsymbol{\Sigma}_{\beta_1}$ indicate the prior mean vector and covariance matrix of $\boldsymbol{\beta}_{1k}$ respectively; $\boldsymbol{X}_k$ is an $N_k \times p_1$ design matrix, where N_k denotes the number of observations in class k; $\boldsymbol{z}_k$ is an $N_k \times 1$ vector of the latent z_{ij} collections and $\boldsymbol{b}_{1k}$ is an $N_k \times 1$ concatenation of the random intercept b_{1i} for class k.

(iv) Denote $\boldsymbol{y}_k^*$ as a $\zeta_k \times 1$ sub-vector of nonzero observations in class k, $\boldsymbol{X}_{2k}^*$ the corresponding $\zeta_k \times p_2$ design matrix and $\boldsymbol{b}_{2k}^*$ a $\zeta_k \times 1$ concatenation of the random intercept b_{2i} restricted to positive observations for class k. For the fixed effect coefficient vector $\boldsymbol{\beta}_{2k}$ of the Kumaraswamy component in model (4.8), we consider a logit link g_κ for modelling quantile κ_{ijk}, that is, $\text{logit}(\kappa_{ijk}) = \boldsymbol{x}_{2ij}^T \boldsymbol{\beta}_{2k} + b_{2i}, k = 1, \ldots, K$. Assuming a multivariate normal prior $\pi(\boldsymbol{\beta}_{2k}) = N_{p_2}(\boldsymbol{\mu}_{\beta_2}, \boldsymbol{\Sigma}_{\beta_2})$, the conditional posterior of $\boldsymbol{\beta}_{2k}$ is obtained by

$$
\begin{aligned}
f(\boldsymbol{\beta}_{2k}|\boldsymbol{y}_k^*, \boldsymbol{b}_{2k}^*, \varphi_k, q) &\propto \Big\{ \prod_{\substack{\forall i,j\, s.t.\, y_{ij}>0 \\ i:C_i=k}} f(y_{ij}|C_i=k, \boldsymbol{\beta}_{2k}, b_{2i}, \varphi_k, q) \Big\} \pi(\boldsymbol{\beta}_{2k}) \\
&= \Big\{ \prod_{\substack{\forall i,j\, s.t.\, y_{ij}>0 \\ i:C_i=k}} \frac{-\log(1-q)\varphi_k}{\log(1-e^{-\varphi_k})\log(\kappa_{ijk})} y_{ij}^{-\frac{\varphi_k}{\log(\kappa_{ijk})}-1} \Big[1 - y_{ij}^{-\frac{\varphi_k}{\log(\kappa_{ijk})}}\Big]^{\frac{\log(1-q)}{\log(1-e^{-\varphi_k})}-1} \Big\} N_{p_2}(\boldsymbol{\mu}_{\beta_2}, \boldsymbol{\Sigma}_{\beta_2}),
\end{aligned}
\tag{4.14}
$$

where $\kappa_{ijk} = e^{\boldsymbol{x}_{2ij}^T \boldsymbol{\beta}_{2k} + b_{2i}} / (1 + e^{\boldsymbol{x}_{2ij}^T \boldsymbol{\beta}_{2k} + b_{2i}})$. Conditional on $C_i = k$, we update $\boldsymbol{\beta}_{2k}$ via a random-walk Metropolis algorithm with a multivariate p_2-dimensional normal proposal centered at previous value $\boldsymbol{\beta}_{2k}^{\text{old}}$ and with covariance matrix $s_{p_2} \boldsymbol{R}_{\beta_{2k}}$ where s_{p_2} is a scaling factor for optimal acceptance rate achievement and $\boldsymbol{R}_{\beta_{2k}}$ is tuned using the empirical covariance obtained from an extended burn-in period of the chain.

(v) Assuming a noninformative inverse gamma prior on the Kumaraswamy precision φ_k, its conditional posterior is then given by

$$
\begin{aligned}
f(\varphi_k|\boldsymbol{y}_k^*, \boldsymbol{\beta}_{2k}, \boldsymbol{b}_{2k}^*, q) &\propto \Big\{ \prod_{\substack{\forall i,j\, s.t.\, y_{ij}>0 \\ i:C_i=k}} f(y_{ij}|C_i=k, \boldsymbol{\beta}_{2k}, b_{2i}, \varphi_k, q) \Big\} \pi(\varphi_k) \\
&= \Big\{ \prod_{\substack{\forall i,j\, s.t.\, y_{ij}>0 \\ i:C_i=k}} \frac{-\log(1-q)\varphi_k}{\log(1-e^{-\varphi_k})\log(\kappa_{ijk})} y_{ij}^{-\frac{\varphi_k}{\log(\kappa_{ijk})}-1} \Big[1 - y_{ij}^{-\frac{\varphi_k}{\log(\kappa_{ijk})}}\Big]^{\frac{\log(1-q)}{\log(1-e^{-\varphi_k})}-1} \Big\} \mathcal{IG}(\rho_1, \rho_2),
\end{aligned}
\tag{4.15}
$$

where $\kappa_{ijk}, \boldsymbol{y}_k^*$ and $\boldsymbol{b}_{2k}^*$ are defined in (iv). Again, we resort to random-walk Metropolis to draw φ_k from its full conditional (4.15). Specifically, we generate and only keep positive candidate values drawn from a normal proposal $N(\varphi_k^{\text{old}}, \sigma_{\varphi_k}^2)$, where φ_k^{old} is the realization of φ_k at previous iteration and $\sigma_{\varphi_k}^2$ is tuned to achieve the acceptance rate around 23%.

(vi) Under an inverse-Wishart prior $\pi(\boldsymbol{\Sigma}_k) = \mathcal{IW}(\tau_0, \boldsymbol{\Psi}_0)$, we update $\boldsymbol{\Sigma}_k$ from its full inverse-Wishart conditional:

$$
\begin{aligned}
f(\boldsymbol{\Sigma}_k|\cdot) &\propto \Big\{ \prod_{i:C_i=k} |\boldsymbol{\Sigma}_k|^{-\frac{1}{2}} e^{\frac{1}{2}\boldsymbol{b}_i^T \boldsymbol{\Sigma}_k^{-1} \boldsymbol{b}_i} \Big\} \pi(\boldsymbol{\Sigma}_k) \\
&\propto |\boldsymbol{\Sigma}_k|^{\frac{m_k}{2}} e^{\frac{1}{2}\mathrm{tr}\left((\sum_{i:C_i=k} \boldsymbol{b}_i \boldsymbol{b}_i^T)\boldsymbol{\Sigma}_k^{-1}\right)} |\boldsymbol{\Sigma}_k|^{-\frac{\tau_0+2+1}{2}} e^{-\frac{1}{2}\mathrm{tr}(\boldsymbol{\Psi}_0 \boldsymbol{\Sigma}_k^{-1})} \\
&= |\boldsymbol{\Sigma}_k|^{\frac{m_k+\tau_0+2+1}{2}} e^{\frac{1}{2}\mathrm{tr}\left((\boldsymbol{\Psi}_0 + \boldsymbol{b}_k^{\dagger\,T} \boldsymbol{b}_k^{\dagger})\boldsymbol{\Sigma}_k^{-1}\right)},
\end{aligned}
\tag{4.16}
$$

which is an $\mathcal{IW}(m_k + \tau_0, \boldsymbol{\Psi}_0 + \boldsymbol{b}_k^{\dagger\,T} \boldsymbol{b}_k^{\dagger})$, where $m_k = \sum_i I(C_i = k)$ is defined in sec. 4.2.2.1 and $\boldsymbol{b}_k^{\dagger}$ is an $m_k \times 2$ matrix with the first column containing b_{1i}'s and the second column containing b_{2i}'s for subject i assigned to the k-th class.

(vii) The full conditional of the stacked random effects vector $\boldsymbol{b}_i = (b_{1i}, b_{2i})^T$ is as follows:

$$
\begin{aligned}
f(\boldsymbol{b}_i|\cdot) &\propto f(\boldsymbol{y}_i|C_i = k, \boldsymbol{\beta}_{1k}, \boldsymbol{\beta}_{2k}, \boldsymbol{b}_i, \varphi_k, q)\pi(\boldsymbol{b}_i|C_i = k) \\
&= f(\boldsymbol{y}_i|C_i = k, \boldsymbol{\beta}_{1k}, \boldsymbol{\beta}_{2k}, \boldsymbol{b}_i, \varphi_k, q) N_2(\boldsymbol{0}, \boldsymbol{\Sigma}_k),
\end{aligned}
\tag{4.17}
$$

where $\boldsymbol{y}_i = (y_{i1}, \ldots, y_{in_i})^T$. Conditional on $c_i = k$, we update $\boldsymbol{b}_i$ via the seminal idea of adaptive random-walk Metropolis algorithm. We consider a bivariate normal proposal $N_2(\boldsymbol{b}_i^{\text{old}}, s_2\boldsymbol{R}_{b_i})$, where $\boldsymbol{b}_i^{\text{old}}$ is the previous iteration value of $\boldsymbol{b}_i$, $\boldsymbol{R}_{b_i}$ is the proposal covariance adjusted according to the empirical covariance obtained from an extended burn-in period and s_2 is the proposal scale based on the optimal acceptance rates.

4.2.2.3 Sampling schemes

Appealing to the presented results, the Bayesian sampling schemes for the posterior computation of the proposed two-part latent class Kumaraswamy quantile mixed model proceed by iterating the following steps after initializing values:

(i) Draw $\phi_k, k = 1, \ldots, K$ using the adaptive random-walk Metropolis algorithm;

(ii) Sample C_i from its full categorical conditional with updated probabilities $\bar{\delta}_{ik}$ given in (4.12);

(iii) Given the current value of $C_i = k, k = 1, \ldots, K$, update the class-specific parameters $\boldsymbol{\beta}_{1k}, \boldsymbol{\beta}_{2k}, \varphi_k$ and $\boldsymbol{\Sigma}_k$:

- (iii.a) Draw $\boldsymbol{\beta}_{1k}$ from its full p_1-dimensional normal conditional formulated in (4.13).
- (iii.b) Draw $\boldsymbol{\beta}_{2k}$ using the adaptive random-walk Metropolis algorithm.
- (iii.c) Draw φ_k by the adaptive random-walk Metropolis algorithm.
- (iii.d) Draw $\boldsymbol{\Sigma}_k$ from its full inverse-Wishart conditional given in (4.16).

(iv) Update $\boldsymbol{b}_i$ using the adaptive random-walk Metropolis algorithm.

A well-known computational issue emerged in Bayesian finite mixture models is the so-called label switching in MCMC outputs, where samples of component-specific parameters may be associated with different class labels during the MCMC run. Consequently, the ergodic averages of component-specific quantities will coincide and become invalid for model inference. Several attempts to solve the label switching have focused on imposing suitable identifiability constraints (Richardson and Green, 1997; Lenk and DeSarbo, 2000; Frühwirth-Schnatter, 2001). However, as Stephens (2000) states, the label switching problem may remain after imposing an identifiability constraint if the constraint has not been carefully chosen. As an alternative, Stephens (2000) proposed a post hoc relabelling algorithm which iteratively minimizes the Kullback-Leibler divergence between an averaged matrix of classification probabilities during the course of the MCMC run, and the classification matrix in each MCMC iteration. We adopt Stephens' approach to address the potential label switching problem before assessing posterior convergence for the proposed model.

4.2.3 Model selection

The DIC for model comparison was developed by Spiegelhalter et al. (2002) and takes the form

$$\begin{aligned} DIC &= \overline{D(\boldsymbol{\Psi})} + p_D \\ &= \overline{D(\boldsymbol{\Psi})} + (\overline{D(\boldsymbol{\Psi})} - D(\widetilde{\boldsymbol{\Psi}})) \\ &= 2\overline{D(\boldsymbol{\Psi})} - D(\widetilde{\boldsymbol{\Psi}}) \\ &= -4\boldsymbol{E}_\theta[\log f(\boldsymbol{y}|\boldsymbol{\Psi})|\boldsymbol{y}] + 2\log f(\boldsymbol{y}|\widetilde{\boldsymbol{\Psi}}), \end{aligned}$$

where $\overline{D(\boldsymbol{\Psi})} = \boldsymbol{E}_\Psi[-2\log f(\boldsymbol{y}|\boldsymbol{\Psi})|\boldsymbol{y}]$ is the posterior mean deviance. p_D is referred to as the effective number of parameters and defined as $(\overline{D(\boldsymbol{\Psi})} - D(\widetilde{\boldsymbol{\Psi}}))$, wherein $\widetilde{\boldsymbol{\Psi}}$ is an estimate of $\boldsymbol{\Psi}$ depending on the data $\boldsymbol{y}$. One commonly employed estimate is the posterior mean $\widetilde{\boldsymbol{\Psi}} = \bar{\boldsymbol{\Psi}} = \boldsymbol{E}[\boldsymbol{\Psi}|\boldsymbol{y}]$ (Gelman et al., 2014).

Celeux et al. (2006) proposed an adapted version of DIC, referred to as DIC_3, which computes $D(\widetilde{\boldsymbol{\Psi}})$ by estimating the posterior mean of the marginal likelihood across all classes, rather than relying on an estimate of the parameters depending on data as in DIC. For mixture models, DIC_3 is preferred because the density of mixtures remains invariant to label switching, rendering this approach more stable compared to utilizing the posterior mean of the parameters. Specifically,

$$\begin{aligned} DIC_3 &= \overline{D(\boldsymbol{\Psi})} + p_{D3} \\ &= \overline{D(\boldsymbol{\Psi})} + (\overline{D(\boldsymbol{\Psi})} - D(\widetilde{\boldsymbol{\Psi}})_3) \\ &= 2\overline{D(\boldsymbol{\Psi})} - D(\widetilde{\boldsymbol{\Psi}})_3 \\ &= -4\boldsymbol{E}_\Psi[\log f(\boldsymbol{y}|\boldsymbol{\Psi})|\boldsymbol{y}] + 2\log \hat{f}(\boldsymbol{y}), \end{aligned}$$

where $D(\widetilde{\boldsymbol{\Psi}})_3 = -2\log\hat{\boldsymbol{y}}$. Both $\overline{D(\boldsymbol{\Psi})}$ and $D(\widetilde{\boldsymbol{\Psi}})_3$ can be approximated using M simulated values $\boldsymbol{\Psi}^{(1)}, \ldots, \boldsymbol{\Psi}^{(M)}$ from MCMC chains. In particular,

$$\begin{aligned} \overline{D(\boldsymbol{\Psi})} &= -2\frac{1}{M}\sum_{m=1}^{M} \log \prod_{i=1}^{N} \sum_{k=1}^{K} \delta_{ik}^{(m)} f(y_{ik}|\boldsymbol{\Psi}_{ik}^{(m)}), \\ D(\widetilde{\boldsymbol{\Psi}})_3 &= -2\log\frac{1}{M}\sum_{m=1}^{M}\prod_{i=1}^{N}\sum_{k=1}^{K} \delta_{ik}^{(m)} f(y_{ik}|\boldsymbol{\Psi}_{ik}^{(m)}). \end{aligned}$$

We employ the DIC_3 as model selection criteria in our real data application analysis.

4.3 Simulation studies

We conduct a simple simulation study to assess the performance of our proposed Bayesian two-part latent class Kumaraswamy quantile mixed model. Models

0.5-Quantile Kumaraswamy Mixed Model				
Class	Model Component	Parameter	True Value	Posterior Mean (SD)
1 (18%)	Binomial	β_{111} (Intercept)	0.25	0.2379 (0.2236)
		β_{112} (Time)	-0.5	-0.4463 (0.0435)
	Kumaraswamy	β_{211} (Intercept)	1.5	1.0583 (0.3966)
		β_{212} (Time)	-0.85	-0.7968 (0.1284)
		φ_1 (Precision)	1	0.8822 (0.1155)
	Covariance	σ^2_{11} (Var$[b_{1i}]$)	2	1.5911 (0.4997)
		σ^2_{12} (Var$[b_{2i}]$)	2	1.6785 (0.6384)
		ϱ_1 (Cov$[b_{1i}, b_{2i}]$)	0.5	0.6814 (0.4498)
2 (20.8%)	Binomial	β_{121} (Intercept)	0.2	0.1085 (0.1375)
		β_{122} (Time)	0.2	0.2040 (0.0210)
	Kumaraswamy	β_{221} (Intercept)	1.25	1.1954 (0.0336)
		β_{222} (Time)	1.25	1.2581 (0.0059)
		φ_2 (Precision)	0.5	0.5086 (0.0279)
	Covariance	σ^2_{21} (Var$[b_{1i}]$)	1	1.0472 (0.2902)
		σ^2_{22} (Var$[b_{2i}]$)	1	0.9203 (0.1624)
		ϱ_2 (Cov$[b_{1i}, b_{2i}]$)	-0.15	-0.0688 (0.1370)
3 (61.2%)	Binomial	β_{131} (Intercept)	-0.45	-0.4003 (0.1002)
		β_{132} (Time)	0.35	0.3261 (0.0152)
	Kumaraswamy	β_{231} (Intercept)	0.3	0.3660 (0.0275)
		β_{232} (Time)	0.3	0.3115 (0.0015)
		φ_3 (Precision)	1.25	1.0414 (0.0276)
	Covariance	σ^2_{31} (Var$[b_{1i}]$)	1.5	1.4599 (0.2208)
		σ^2_{32} (Var$[b_{2i}]$)	0.5	0.6361 (0.0620)
		ϱ_3 (Cov$[b_{1i}, b_{2i}]$)	0	0.0976 (0.0740)

SD: Posterior standard deviation.
True class proportions are 17.6%, 22.4% and 60%.

Table 4.1 Posterior summaries for the simulated two-part three-class 0.5-quantile Kumaraswamy mixed model

0.75-Quantile Kumaraswamy Mixed Model				
Class	Model Component	Parameter	True Value	Posterior Mean (SD)
1 (18%)	Binomial	β_{111} (Intercept)	0.25	0.1656 (0.1973)
		β_{112} (Time)	-0.5	-0.4487 (0.0474)
	Kumaraswamy	β_{211} (Intercept)	1.5	1.0928 (0.3210)
		β_{212} (Time)	-0.85	-0.8004 (0.1019)
		φ_1 (Precision)	1	0.8819 (0.1378)
	Covariance	σ^2_{11} (Var[b_{1i}])	2	1.6339 (0.4893)
		σ^2_{12} (Var[b_{2i}])	2	1.7680 (0.6406)
		ϱ_1 (Cov[b_{1i}, b_{2i}])	0.5	0.6832 (0.4351)
2 (21%)	Binomial	β_{121} (Intercept)	0.2	0.1421 (0.1399)
		β_{122} (Time)	0.2	0.2040 (0.0211)
	Kumaraswamy	β_{221} (Intercept)	1.25	1.3319 (0.0856)
		β_{222} (Time)	1.25	1.2522 (0.0161)
		φ_2 (Precision)	0.5	0.5037 (0.0294)
	Covariance	σ^2_{21} (Var[b_{1i}])	1	1.0277 (0.2814)
		σ^2_{22} (Var[b_{2i}])	1	0.9224 (0.1469)
		ϱ_2 (Cov[b_{1i}, b_{2i}])	-0.15	-0.0841 (0.1186)
3 (61%)	Binomial	β_{131} (Intercept)	-0.45	-0.4185 (0.0952)
		β_{132} (Time)	0.35	0.3267 (0.0150)
	Kumaraswamy	β_{231} (Intercept)	0.3	0.2140 (0.0454)
		β_{232} (Time)	0.3	0.3184 (0.0040)
		φ_3 (Precision)	1.25	0.8962 (0.0245)
	Covariance	σ^2_{31} (Var[b_{1i}])	1.5	1.4446 (0.1933)
		σ^2_{32} (Var[b_{2i}])	0.5	0.6548 (0.0648)
		ϱ_3 (Cov[b_{1i}, b_{2i}])	0	0.1241 (0.0769)

SD: Posterior standard deviation.
True class proportions are 17.6%, 22.4% and 60%.

Table 4.2 Posterior summaries for the simulated two-part three-class 0.75-quantile Kumaraswamy mixed model

0.95-Quantile Kumaraswamy Mixed Model

Class	Model Component	Parameter	True Value	Posterior Mean (SD)
1 (18%)	Binomial	β_{111} (Intercept)	0.25	0.1844 (0.2088)
		β_{112} (Time)	-0.5	-0.4392 (0.0422)
	Kumaraswamy	β_{211} (Intercept)	1.5	1.3028 (0.2871)
		β_{212} (Time)	-0.85	-0.8171 (0.0580)
		φ_1 (Precision)	1	0.8781 (0.1447)
	Covariance	σ_{11}^2 (Var$[b_{1i}]$)	2	1.5339 (0.4495)
		σ_{12}^2 (Var$[b_{2i}]$)	2	1.7608 (0.4735)
		ϱ_1 (Cov$[b_{1i}, b_{2i}]$)	0.5	0.5306 (0.3860)
2 (21.2%)	Binomial	β_{121} (Intercept)	0.2	0.1925 (0.1473)
		β_{122} (Time)	0.2	0.2049 (0.0211)
	Kumaraswamy	β_{221} (Intercept)	1.25	1.2639 (0.0001)
		β_{222} (Time)	1.25	1.2488 (0.0001)
		φ_2 (Precision)	0.5	0.5155 (0.0338)
	Covariance	σ_{21}^2 (Var$[b_{1i}]$)	1	1.0904 (0.3017)
		σ_{22}^2 (Var$[b_{2i}]$)	1	0.9915 (0.1434)
		ϱ_2 (Cov$[b_{1i}, b_{2i}]$)	-0.15	-0.1754 (0.1264)
3 (60.8%)	Binomial	β_{131} (Intercept)	-0.45	-0.4640 (0.1020)
		β_{132} (Time)	0.35	0.3255 (0.0156)
	Kumaraswamy	β_{231} (Intercept)	0.3	0.3074 (0.0299)
		β_{232} (Time)	0.3	0.3001 (0.0046)
		φ_3 (Precision)	1.25	0.7119 (0.0245)
	Covariance	σ_{31}^2 (Var$[b_{1i}]$)	1.5	1.4463 (0.2260)
		σ_{32}^2 (Var$[b_{2i}]$)	0.5	0.7740 (0.0676)
		ϱ_3 (Cov$[b_{1i}, b_{2i}]$)	0	0.1097 (0.0818)

SD: Posterior standard deviation.
True class proportions are 17.6%, 22.4% and 60%.

Table 4.3 Posterior summaries for the simulated two-part three-class 0.95-quantile Kumaraswamy mixed model

with different classes were fitted in R (R Core Team, 2021) using the MCMC sampling schemes presented in Section 4.2.2.3. The R code was developed by the authors with some adaptions based upon the code provided in Neelon et al. (2015). We choose three quantile levels: $q = 0.5$, $q = 0.75$ and $q = 0.95$. For the data generation process in each quantile specification scenario, we draw 100 datasets from a three-class model according to formulation (4.8). We consider 500 subjects ($i = 500$), each with 10 observations ($j = 10$). We construct fixed effect covariates matrix $\boldsymbol{T}_{500\times 10}$, with each row consisting of the time indexes $1, 2, \ldots, 10$. The fixed effect covariates for both the binomial and the Kumaraswamy components are then set as $\boldsymbol{x}_{lij} = (x_{lij1}, x_{lij2})^T, l = 1, 2$, where $x_{lij1} = 1$ corresponds to the intercept and x_{lij2} is chosen as class-specific fixed effect time covariates. The parameters are specified as follows: $\boldsymbol{\beta}_{11} = (\beta_{111}, \beta_{112})^T = (0.25, -0.5)^T$, $\boldsymbol{\beta}_{21} = (\beta_{211}, \beta_{212})^T = (1.5, -0.85)^T$, $\varphi_1 = 1$ for class 1; $\boldsymbol{\beta}_{12} = (\beta_{121}, \beta_{122})^T = (0.2, 0.2)^T$, $\boldsymbol{\beta}_{22} = (\beta_{221}, \beta_{222})^T = (1.25, 1.25)^T$, $\varphi_2 = 0.5$ for class 2; $\boldsymbol{\beta}_{13} = (\beta_{131}, \beta_{132})^T = (-0.45, 0.35)^T$, $\boldsymbol{\beta}_{23} = (\beta_{231}, \beta_{232})^T = (0.3, 0.3)^T$, $\varphi_3 = 1.25$ for class 3. Given $C_i = k$, the stacked random effect vectors are set as $\boldsymbol{b}_i = (b_{1i}, b_{2i})^T \sim N_2(\boldsymbol{0}, \boldsymbol{\Sigma}_k), k = 1, 2, 3$, where $\boldsymbol{\Sigma}_1 = \big(\begin{smallmatrix} \sigma^2_{11} & \varrho_1 \\ \varrho_1 & \sigma^2_{12} \end{smallmatrix}\big) = \big(\begin{smallmatrix} 2 & 0.5 \\ 0.5 & 2 \end{smallmatrix}\big)$, $\boldsymbol{\Sigma}_2 = \big(\begin{smallmatrix} \sigma^2_{21} & \varrho_2 \\ \varrho_2 & \sigma^2_{22} \end{smallmatrix}\big) = \big(\begin{smallmatrix} 1 & -0.15 \\ -0.15 & 1 \end{smallmatrix}\big)$ and $\boldsymbol{\Sigma}_3 = \big(\begin{smallmatrix} \sigma^2_{31} & \varrho_3 \\ \varrho_3 & \sigma^2_{32} \end{smallmatrix}\big) = \big(\begin{smallmatrix} 1.5 & 0 \\ 0 & 0.5 \end{smallmatrix}\big)$. Following (4.9), we allow the categorical probability δ_{ik} to be associated with the covariate vector $\boldsymbol{\eta}_i$ which consists of an intercept and independent Bernoulli Ber(0.65) variables. We set $\boldsymbol{\phi}_2 = (0.1, 0.1)^T$, $\boldsymbol{\phi}_3 = (1.75, -0.9)^T$ and $\boldsymbol{\phi}_1 = \boldsymbol{0}$ for identifiability. We assign non-informative priors for formulated model parameters. Specifically, we have $\pi(\boldsymbol{\phi}_k) = \pi(\boldsymbol{\beta}_{1k}) = \pi(\boldsymbol{\beta}_{2k}) = N_2(\boldsymbol{0}, 100\boldsymbol{I}_2)$, $\pi(\varphi_k) = \mathcal{IG}(0.001, 0.001)$ and $\pi(\boldsymbol{\Sigma}_k) = \mathcal{IW}(2, \text{diag}(2))$. For each replication, we discard the first 10,000 iterations of the MCMC chain as a burn-in and run 10,000 iterations with a thinning equal to 5, leading to 2000 samples upon which the posterior inference for each parameter is performed. The CPU time takes around 25 min to complete parameter estimation in each quantile calibration for one simulated dataset.

With the three different quantile level specifications, Tables 4.1-4.3 present the model fitting summaries averaged over the 100 simulations respectively. The first column reports the estimated class percentages, which present a very good approximation to the true class proportions of 17.6%, 22.4% and 60% for class 1-3 separately. The fifth column provides the averaged posterior means with the posterior standard deviations for each parameter of interest. We observe that the bias is reasonably low for all parameters, including the estimation for the random effect covariance components and the estimates under the upper quantile level ($p = 0.95$) scenario. This simulation study indicates our proposed Bayesian approach for the two-part latent class Kumaraswamy quantile mixed model achieves desirable estimation results for regimes covering different quantile levels including extremes.

4.4 Real data analysis

In this section we present the application of the proposed methodology to the RAND Health Insurance Experiment (HIE) dataset. The RAND HIE is one of the longest running and largest controlled health insurance experiments ever conducted in the United States, to assess how the patient's use of health services and quality of care are affected by different types of randomly assigned health insurance. Because of the random assignment, the reliability of insurance coverage and the availability of important health care utilisation variables, the data is widely regarded as the most reliable basis for health insurance and medical demand modelling. The experiment, conducted by the RAND Corporation from 1974 to 1982, collects data from about 8,000 enrollees in 2,823 families from six sites across the USA. Each family was enrolled in one of 14 different HIS insurance plans for either 3 or 5 years. For a full description and discussion of the data, see, e.g., Aron-Dine et al. (2013), Deb and Trivedi (2002), Duan et al. (1983), and Manning et al. (1987). The data is well documented and publicly available in the R package *sampleSelection*.

Our objective is to explore the impact of various covariates, including temporal trends and economic budget characteristics, on the ratio of outpatient expenses (outpdol) to the overall expenditure on health services (meddol) within US households across various quantile levels. Additionally, we aim to uncover latent subgroups within the population. Outpatient expenses encompass all covered outpatient medical services, excluding dental care, outpatient psychotherapy and outpatient drugs or supplies. Total medical expenses comprise all covered inpatient and outpatient services, inclusive of drugs, supplies and inpatient costs for newborns, but excluding dental care and outpatient psychotherapy. We consider an available subsample $N = 4080$ from the complete cases of families who were enrolled in the HIS insurance plans from year 1 to year 5 of the study, where the ratio of outpatient to medical expenditures falls within the range [0,1). We allow the natural logarithm of medical expenses (lnmeddol) to serve as class-membership covariates; the $\boldsymbol{\eta}_i$ in (4.9) therefrom represents a 2×1 vector consisting of an intercept and the considered continuous variables. The fixed effect covariates $\boldsymbol{x}_{ij}$ for both the binomial and the Kumaraswamy components are composed of an intercept term and five dummy indicators representing years 1-5. Here categorical modelling of time is chosen to accommodate optimal flexibility in capturing the time trend. Alternative parameterizations, such as polynomials or splines of the time, may be preferable in scenarios with numerous time points. The summary statistics for the response variable are reported in Table 4.4. From the table, there are 23.92% enrollees who did not have any outpatient expense over the total medical expenditure. Furthermore, the median, 0.25th and 0.75th quantiles of the proportional outpatient expenditure are 0.4412, 0.0216 and 0.7444 respectively, with the maximum proportion being 0.9981. The displayed characteristics indicate the desirability of adopting a two-part quantile regression model for accommodating the data. Specifically,

Mean	SD	Kurtosis	No Exp. (%)	Maximum	q-th quantile		
					0.25	0.50	0.75
0.4247	0.3408	1.5199	23.9216	0.9981	0.0216	0.4412	0.7444

Table 4.4 Summary statistics for the proportion of outpatient expenses to the total medical expenditures

the binomial process will assess the association between the covariates and the probability of having any outpatient expense. The Kumaraswamy process will investigate whether the covariates exert different influences on the proportion of outpatient outgo and provide insights for significant determinants across low and high outpatient cost, such as proportions at the 25th and 75th percentiles.

We implement a series of two-part latent class Kumaraswamy quantile mixed models as in (4.8), allowing the number of classes K to range from two to four. Within each class, we fit mixed effect models with random intercepts for both components under three quantile level scenarios: $q = 0.25$, $q = 0.5$ and $q = 0.75$. We impose diffuse priors for model parameters in each class: $\pi(\boldsymbol{\phi}_k) = N_3(\mathbf{0}, 100\boldsymbol{I}_3)$, $\pi(\boldsymbol{\beta}_{1k}) = \pi(\boldsymbol{\beta}_{2k}) = N_4(\mathbf{0}, 100\boldsymbol{I}_4)$, $\pi(\varphi_k) = \mathcal{IG}(0.001, 0.001)$ and $\pi(\boldsymbol{\Sigma}_k) = \mathcal{IW}(2, \text{diag}(2))$. We run 200,000 iterations for each model, discarding the first 50,000 for the burn-in. We retain every 50th draw for thinning to reduce autocorrelation. We examine trace plots and autocorrelation function plots for all model parameters to assess convergence of the chains. All trace plots showed good overlapping and mixing trajectories, indicating evidence of convergence to the stationary distribution. Little evidence of label switching has emerged within individual chains in our MCMC estimates.

Model comparison is performed by employing DIC_3 measures illuminated in Section 4.2.3. The results are presented in Table 4.5. The DIC_3 kept decreasing as the number of classes increased for each specified quantile level of the proportional outpatient expenditures data. Overall, the four-class Kumaraswamy quantile mixed effects models were preferred, followed by the three- and two-class mixed models. We select the two-part four-class Kumaraswamy quantile mixed regression as our working model based on its lowest information criteria.

Tables 4.6-4.8 provide the posterior means and the 95% highest posterior density (HPD) intervals for the four-class models at $q = 0.25$, $q = 0.5$ and $q = 0.75$ respectively. For the fitted 0.25-quantile Kumaraswamy mixed model, the first class comprised an estimated 34.19% of the population and we term this group as “high-first quartile spenders”. Participants from this class had a relatively high initial probability of outpatient spending and then a rapid decreasing trend until year 4. The level of the 25th percentile spending was high in years 1-3 and then decreased until year 5. Class 2 included 36.76% of the participants and the trajectory pattern was characterized by a low probability of initial spending and prudent first quartile spending. For these individuals, a high

Model Description	Number of Classes	p_{D3}	DIC_3
0.25-Quantile Kumaraswamy Mixed Regression	2	509.733	-1603.395
	3	348.176	-5097.910
	4	438.688	**-7562.129**
0.5-Quantile Kumaraswamy Mixed Regression	2	516.127	-1610.492
	3	374.535	-5217.575
	4	440.337	**-7471.915**
0.75-Quantile Kumaraswamy Mixed Regression	2	543.170	-1546.525
	3	380.942	-5217.156
	4	408.757	**-7597.464**

Table 4.5 Model comparison statistics for the proportional outpatient expenses data

probability of spending during years 2-5 but an average spending level around 0 during years 2-4 with a rapid decreasing trend in year 5 could be observed. Enrollees from this class were labelled as "light-first quartile spenders". The third class embraced 16.79% of participants and we call this group "heavy-first quartile spenders". Individuals from this class were characterized by a high initial probability of spending and high 25th percentile of proportional outpatient expenses. In this scenario, except for year 3, all the other four time indicators showed non-negligible (i.e., the 95% HPD interval for the regression coefficient does not include 0) positive relationships with the spending, with year 4 and year 2 being the most remarkable factors followed by year 1 and year 5. Class 4 incorporated 12.26% of subjects and showed a comparatively high probability of spending during years 3-4 but a very low level of the 25th percentile of proportional spending. These participants, defined as "non/occasional-first quartile spenders", were relatively rare users of outpatient medical services who exhibited a spending pattern occasionally at a very low level in year 3.

For the formulated median Kumaraswamy mixed model, the first class comprised 43.26% of subjects and exhibited a comparatively high initial probability of spending along with median expenditure proportions. Notably, all the five time indicators were significant and positively associated with the 50th percentile of outpatient expense proportions, with year 1 and year 3 being the most influential time factors. We designate this class as "high-median spenders". The second class, encompassing an estimated 32.11% of participants, demonstrated a low initial probability of spending and an overall light level of median proportional outpatient expenditures. These individuals displayed a significant expenditure on outpatient medical services in year 1, followed by diminishing costs in years 2-4, with a declining trend in year 5. Participants in this class are characterized as "light-median spenders". Embracing 13.36% of participants, class 3, labelled as "non/occasional-median spenders", showed a high initial probability of spending and high baseline median proportional outpatient expenses. How-

Class (%)	Model Component	Parameter (Covariate)	Posterior Mean	95% HPD Interval
1 (34.19%)	Binomial	β_{111} (Intercept)	1.1357	(-0.1875, 2.3976)
		β_{112} (Year$_1$)	2.0759	(0.7346, 3.6103)
		β_{113} (Year$_2$)	-0.0666	(-1.3319, 1.2608)
		β_{114} (Year$_3$)	-0.3339	(-1.6723, 0.9489)
		β_{115} (Year$_4$)	-0.1864	(-1.5416, 1.1200)
		β_{116} (Year$_5$)	-0.2321	(-1.4662, 1.1811)
	Kumaraswamy	β_{211} (Intercept)	-0.8447	(-1.4150, -0.1859)
		β_{212} (Year$_1$)	0.6583	(0.0530, 1.2466)
		β_{213} (Year$_2$)	0.4041	(-0.1960, 1.0244)
		β_{214} (Year$_3$)	0.5451	(0.0656, 1.0248)
		β_{215} (Year$_4$)	0.1201	(-0.3133, 0.8402)
		β_{216} (Year$_5$)	0.2366	(-0.4015, 0.7751)
		φ_1 (Precision)	1.4309	(1.3378, 1.5132)
	Covariance	σ^2_{11} (Var$[b_{1i}]$)	0.8095	(0.4883, 1.1633)
		σ^2_{12} (Var$[b_{2i}]$)	0.3278	(0.1445, 0.4820)
		ϱ_1 (Cov$[b_{1i}, b_{2i}]$)	0.1946	(0.0372, 0.3561)
2 (36.76%)	Binomial	β_{121} (Intercept)	-1.0701	(-2.3142, 0.2786)
		β_{122} (Year$_1$)	-3.3467	(-5.2086, -1.7352)
		β_{123} (Year$_2$)	0.4911	(-0.9351, 1.6125)
		β_{124} (Year$_3$)	0.3741	(-0.9554, 1.6172)
		β_{125} (Year$_4$)	0.3992	(-0.9439, 1.6441)
		β_{126} (Year$_5$)	0.6971	(-0.6652, 1.9477)
	Kumaraswamy	β_{221} (Intercept)	-0.5803	(-1.8355, 0.4908)
		β_{222} (Year$_1$)	0.2307	(-1.7838, 2.4397)
		β_{223} (Year$_2$)	-0.0582	(-1.1445, 1.1134)
		β_{224} (Year$_3$)	0.0599	(-1.0557, 1.1918)
		β_{225} (Year$_4$)	0.0471	(-1.1459, 1.4068)
		β_{226} (Year$_5$)	-0.2160	(-1.3691, 0.9850)
		φ_2 (Precision)	1.4237	(1.2404, 1.6317)
	Covariance	σ^2_{21} (Var$[b_{1i}]$)	1.3510	(0.7121, 1.9853)
		σ^2_{22} (Var$[b_{2i}]$)	1.0082	(0.1992, 1.8279)
		ϱ_2 (Cov$[b_{1i}, b_{2i}]$)	-0.1263	(-0.8069, 0.5340)
3 (16.79%)	Binomial	β_{131} (Intercept)	2.9787	(1.2556, 4.7502)
		β_{132} (Year$_1$)	0.4620	(-1.7143, 3.1251)
		β_{133} (Year$_2$)	-0.2092	(-2.2136, 2.1166)
		β_{134} (Year$_3$)	1.1081	(-1.2103, 3.4454)
		β_{135} (Year$_4$)	0.8899	(-1.1381, 2.9818)
		β_{136} (Year$_5$)	0.7161	(-1.1716, 3.2447)
	Kumaraswamy	β_{231} (Intercept)	-0.7653	(-0.8960, -0.6299)
		β_{232} (Year$_1$)	0.4058	(0.0748, 0.6197)
		β_{233} (Year$_2$)	0.5283	(0.3192, 0.7446)
		β_{234} (Year$_3$)	0.1095	(-0.2101, 0.3902)
		β_{235} (Year$_4$)	0.5352	(0.2919, 0.7091)
		β_{236} (Year$_5$)	0.2391	(0.0180, 0.5082)
		φ_3 (Precision)	2.4138	(2.1461, 2.6805)
	Covariance	σ^2_{31} (Var$[b_{1i}]$)	0.6362	(0.1286, 1.6511)
		σ^2_{32} (Var$[b_{2i}]$)	0.7161	(0.4601, 0.9970)
		ϱ_3 (Cov$[b_{1i}, b_{2i}]$)	-0.0115	(-0.5812, 0.6327)
4 (12.26%)	Binomial	β_{141} (Intercept)	2.1137	(0.8631, 3.5185)
		β_{142} (Year$_1$)	1.0454	(-0.3461, 2.5421)
		β_{143} (Year$_2$)	0.0202	(-1.3919, 1.2145)
		β_{144} (Year$_3$)	0.3866	(-0.9391, 1.7084)
		β_{145} (Year$_4$)	0.5609	(-0.9418, 1.8283)
		β_{146} (Year$_5$)	0.1381	(-1.2533, 1.4339)
	Kumaraswamy	β_{241} (Intercept)	-1.2702	(-2.5155, 0.0992)
		β_{242} (Year$_1$)	-1.7083	(-3.1902, -0.4547)
		β_{243} (Year$_2$)	-0.3462	(-2.1322, 0.7296)
		β_{244} (Year$_3$)	0.0391	(-1.5053, 1.4051)
		β_{245} (Year$_4$)	-0.0301	(-1.3065, 1.1697)
		β_{246} (Year$_5$)	-0.0605	(-1.4098, 1.1652)
		φ_4 (Precision)	1.5004	(1.3831, 1.6490)
	Covariance	σ^2_{41} (Var$[b_{1i}]$)	2.3814	(0.9063, 4.4555)
		σ^2_{42} (Var$[b_{2i}]$)	0.6366	(0.1881, 1.1457)
		ϱ_4 (Cov$[b_{1i}, b_{2i}]$)	0.7374	(0.1432, 1.7306)

Table 4.6 Posterior means and 95% HPD intervals for the two-part four-class 0.25-quantile Kumaraswamy mixed model

Class (%)	Model Component	Parameter (Covariate)	Posterior Mean	95% HPD Interval
1 (43.26%)	Binomial	β_{111} (Intercept)	1.3213	(-0.0082, 2.5861)
		β_{112} (Year$_1$)	2.3489	(0.6688, 4.6767)
		β_{113} (Year$_2$)	-0.1206	(-1.4035, 1.1494)
		β_{114} (Year$_3$)	-0.3443	(-1.5207, 1.0090)
		β_{115} (Year$_4$)	-0.2129	(-1.5062, 1.0842)
		β_{116} (Year$_5$)	-0.2549	(-1.4855, 1.0163)
	Kumaraswamy	β_{211} (Intercept)	-0.0317	(-0.1535, 0.0071)
		β_{212} (Year$_1$)	0.7590	(0.7270, 0.8914)
		β_{213} (Year$_2$)	0.5951	(0.5835, 0.6447)
		β_{214} (Year$_3$)	0.6935	(0.6336, 0.8590)
		β_{215} (Year$_4$)	0.4532	(0.4196, 0.5559)
		β_{216} (Year$_5$)	0.6006	(0.5816, 0.6603)
		φ_1 (Precision)	0.7807	(0.7275, 0.8368)
	Covariance	σ^2_{11} (Var[b_{1i}])	0.9242	(0.6208, 1.3106)
		σ^2_{12} (Var[b_{2i}])	0.2689	(0.1771, 0.3674)
		ϱ_1 (Cov[b_{1i}, b_{2i}])	0.1990	(0.0427, 0.3332)
2 (32.11%)	Binomial	β_{121} (Intercept)	-1.0617	(-2.2584, 0.3595)
		β_{122} (Year$_1$)	-3.2133	(-5.0556, -1.2381)
		β_{123} (Year$_2$)	0.4808	(-0.9967, 1.6495)
		β_{124} (Year$_3$)	0.3795	(-0.9972, 1.6282)
		β_{125} (Year$_4$)	0.3843	(-0.9493, 1.7718)
		β_{126} (Year$_5$)	0.6803	(-0.6630, 2.0175)
	Kumaraswamy	β_{221} (Intercept)	0.3245	(-1.1245, 1.6804)
		β_{222} (Year$_1$)	0.8014	(-1.6515, 3.3192)
		β_{223} (Year$_2$)	0.0162	(-1.4925, 1.4539)
		β_{224} (Year$_3$)	0.1034	(-1.1822, 1.6228)
		β_{225} (Year$_4$)	0.1179	(-1.3451, 1.5539)
		β_{226} (Year$_5$)	-0.0993	(-1.5291, 1.3514)
		φ_2 (Precision)	0.7329	(0.5902, 0.9044)
	Covariance	σ^2_{21} (Var[b_{1i}])	1.3477	(0.7148, 1.9810)
		σ^2_{22} (Var[b_{2i}])	0.6441	(0.1628, 1.1821)
		ϱ_2 (Cov[b_{1i}, b_{2i}])	-0.0767	(-0.5888, 0.4592)
3 (13.36%)	Binomial	β_{131} (Intercept)	2.9674	(1.6054, 4.9743)
		β_{132} (Year$_1$)	-0.0940	(-2.0541, 1.6595)
		β_{133} (Year$_2$)	0.3589	(-1.7405, 2.8990)
		β_{134} (Year$_3$)	1.0029	(-0.8776, 3.2685)
		β_{135} (Year$_4$)	1.0906	(-1.4470, 3.4211)
		β_{136} (Year$_5$)	0.8206	(-1.3554, 2.8342)
	Kumaraswamy	β_{231} (Intercept)	0.7305	(0.5800, 1.0466)
		β_{232} (Year$_1$)	-0.6323	(-1.0567, -0.4435)
		β_{233} (Year$_2$)	-0.6929	(-1.0116, -0.5324)
		β_{234} (Year$_3$)	-1.0934	(-1.4977, -0.8822)
		β_{235} (Year$_4$)	-0.7983	(-1.1643, -0.5846)
		β_{236} (Year$_5$)	-1.2199	(-1.4397, -0.9837)
		φ_3 (Precision)	1.7698	(1.4929, 2.0295)
	Covariance	σ^2_{31} (Var[b_{1i}])	0.9045	(0.0917, 3.3553)
		σ^2_{32} (Var[b_{2i}])	0.4222	(0.1967, 0.6758)
		ϱ_3 (Cov[b_{1i}, b_{2i}])	0.1688	(-0.3614, 1.1215)
4 (11.27%)	Binomial	β_{141} (Intercept)	2.2175	(0.6901, 3.5672)
		β_{142} (Year$_1$)	1.2597	(-0.5648, 3.1138)
		β_{143} (Year$_2$)	-0.0431	(-1.1999, 1.3319)
		β_{144} (Year$_3$)	0.3687	(-0.8487, 1.6506)
		β_{145} (Year$_4$)	0.5582	(-0.6327, 1.9989)
		β_{146} (Year$_5$)	0.0876	(-1.1249, 1.3714)
	Kumaraswamy	β_{241} (Intercept)	-1.0028	(-1.7131, -0.3775)
		β_{242} (Year$_1$)	-0.3499	(-1.1639, 0.5301)
		β_{243} (Year$_2$)	0.5464	(-0.0741, 1.3153)
		β_{244} (Year$_3$)	0.9049	(0.3365, 1.5666)
		β_{245} (Year$_4$)	0.8426	(0.1589, 1.5486)
		β_{246} (Year$_5$)	0.8612	(0.2265, 1.5821)
		φ_4 (Precision)	0.7641	(0.6531, 0.8558)
	Covariance	σ^2_{41} (Var[b_{1i}])	2.5543	(0.7660, 4.7494)
		σ^2_{42} (Var[b_{2i}])	0.3256	(0.0985, 0.5900)
		ϱ_4 (Cov[b_{1i}, b_{2i}])	0.4912	(-0.0214, 1.1461)

Table 4.7 Posterior means and 95% HPD intervals for the two-part four-class 0.5-quantile Kumaraswamy mixed model

Class (%)	Model Component	Parameter (Covariate)	Posterior Mean	95% HPD Interval
1 (14.34%)	Binomial	β_{111} (Intercept)	2.9441	(1.3846, 4.8740)
		β_{112} (Year$_1$)	0.0728	(-2.1435, 1.9469)
		β_{113} (Year$_2$)	0.1398	(-1.6477, 2.4935)
		β_{114} (Year$_3$)	0.9612	(-0.9428, 3.3729)
		β_{115} (Year$_4$)	0.7499	(-1.7570, 2.8529)
		β_{116} (Year$_5$)	0.7142	(-0.8877, 2.8719)
	Kumaraswamy	β_{211} (Intercept)	1.0527	(0.8258, 1.2605)
		β_{212} (Year$_1$)	-0.1081	(-0.3692, 0.1688)
		β_{213} (Year$_2$)	0.0062	(-0.2795, 0.1872)
		β_{214} (Year$_3$)	-0.2346	(-0.4888, -0.0028)
		β_{215} (Year$_4$)	-0.1214	(-0.5228, 0.2710)
		β_{216} (Year$_5$)	-0.2742	(-0.6303, 0.0581)
		φ_1 (Precision)	0.9063	(0.7475, 1.0680)
	Covariance	σ^2_{11} (Var$[b_{1i}]$)	0.5839	(0.0837, 1.6199)
		σ^2_{12} (Var$[b_{2i}]$)	0.4283	(0.2462, 0.5864)
		ϱ_1 (Cov$[b_{1i}, b_{2i}]$)	0.0704	(-0.3210, 0.5146)
2 (32.84%)	Binomial	β_{121} (Intercept)	2.1698	(0.9328, 3.5042)
		β_{122} (Year$_1$)	1.1089	(-0.2819, 2.4520)
		β_{123} (Year$_2$)	0.0179	(-1.1057, 1.2042)
		β_{124} (Year$_3$)	0.3824	(-0.8735, 1.5923)
		β_{125} (Year$_4$)	0.5805	(-0.7514, 1.8451)
		β_{126} (Year$_5$)	0.1703	(-1.0621, 1.4777)
	Kumaraswamy	β_{221} (Intercept)	1.2827	(0.6884, 1.9102)
		β_{222} (Year$_1$)	-1.2246	(-1.8244, -0.7369)
		β_{223} (Year$_2$)	-0.5846	(-1.3994, -0.1002)
		β_{224} (Year$_3$)	-0.3709	(-1.0432, 0.0678)
		β_{225} (Year$_4$)	-0.3666	(-0.8412, 0.3145)
		β_{226} (Year$_5$)	-0.3541	(-0.9863, 0.0207)
		φ_2 (Precision)	0.3399	(0.2871, 0.3980)
	Covariance	σ^2_{21} (Var$[b_{1i}]$)	0.5839	(0.0837, 1.6199)
		σ^2_{22} (Var$[b_{2i}]$)	0.4283	(0.2462, 0.5864)
		ϱ_2 (Cov$[b_{1i}, b_{2i}]$)	0.0704	(-0.3210, 0.5146)
3 (33.82%)	Binomial	β_{131} (Intercept)	1.1156	(-0.0925, 2.3173)
		β_{132} (Year$_1$)	2.0520	(0.6488, 3.5159)
		β_{133} (Year$_2$)	-0.1062	(-1.2602, 1.1446)
		β_{134} (Year$_3$)	-0.3573	(-1.4756, 0.9439)
		β_{135} (Year$_4$)	-0.2163	(-1.4603, 0.9116)
		β_{136} (Year$_5$)	-0.2569	(-1.4752, 0.9116)
	Kumaraswamy	β_{231} (Intercept)	0.6328	(0.5449, 0.7722)
		β_{232} (Year$_1$)	1.0552	(0.9167, 1.1768)
		β_{233} (Year$_2$)	0.8236	(0.6552, 0.9488)
		β_{234} (Year$_3$)	0.9064	(0.7996, 0.9918)
		β_{235} (Year$_4$)	0.7894	(0.5719, 0.9528)
		β_{236} (Year$_5$)	0.7938	(0.7199, 0.9564)
		φ_3 (Precision)	0.3168	(0.2808, 0.3564)
	Covariance	σ^2_{31} (Var$[b_{1i}]$)	0.7382	(0.4493, 1.0416)
		σ^2_{32} (Var$[b_{2i}]$)	0.1908	(0.1111, 0.2734)
		ϱ_3 (Cov$[b_{1i}, b_{2i}]$)	0.1320	(-0.0021, 0.2412)
4 (19.00%)	Binomial	β_{141} (Intercept)	-1.1212	(-2.4012, 0.2125)
		β_{142} (Year$_1$)	-3.1880	(-5.0173, -1.3034)
		β_{143} (Year$_2$)	0.5430	(-0.8488, 1.7900)
		β_{144} (Year$_3$)	0.4375	(-0.8454, 1.7640)
		β_{145} (Year$_4$)	0.4531	(-0.9063, 1.6596)
		β_{146} (Year$_5$)	0.7639	(-0.5162, 2.0945)
	Kumaraswamy	β_{241} (Intercept)	1.2207	(0.1343, 2.3102)
		β_{242} (Year$_1$)	-1.2629	(-4.0363, 1.9640)
		β_{243} (Year$_2$)	0.2226	(-0.8911, 1.2510)
		β_{244} (Year$_3$)	0.3363	(-0.6203, 1.5274)
		β_{245} (Year$_4$)	0.3071	(-0.6281, 1.7350)
		β_{246} (Year$_5$)	0.0988	(-0.9122, 1.3839)
		φ_4 (Precision)	0.3113	(0.2131, 0.4176)
	Covariance	σ^2_{41} (Var$[b_{1i}]$)	1.3211	(0.7230, 1.9126)
		σ^2_{42} (Var$[b_{2i}]$)	0.4977	(0.1225, 0.8824)
		ϱ_4 (Cov$[b_{1i}, b_{2i}]$)	-0.1221	(-0.6012, 0.2809)

Table 4.8 Posterior means and 95% HPD intervals for the two-part four-class 0.75-quantile Kumaraswamy mixed model

ever, their median proportional spending remained very low throughout years 1-5. All the five time indicators exhibited non-negligible negative relationships with spending, with year 3 and year 5 standing out as the most prominent factors, followed by year 4. Only 11.27% of participants fell into class 4, described as "heavy-median spenders". This class demonstrated a high probability of spending and substantial median spending. Individuals in this group were characterized as chronic users of outpatient medical services with consistently high expenditure trends, particularly evident throughout years 2-5.

The first class in the assessed Kumaraswamy mixed model at the third quartile comprised 14.34% of participants and was termed "light-third quartile spenders". Individuals in this category exhibited a high initial probability of spending but consistently low 75th percentile of proportional outpatient expenses. Notably, Year 3 and year 5 exerted remarkable negative influence on this upper level of spending. Class 2, encompassing an estimated 32.84% of participants, displayed a high initial probability of spending alongside baseline 0.75-quantile proportional outpatient expenses. However, their upper level of proportional spending remained very low throughout years 1-5. All the five time indicators exhibited negative relationships with spending. These subjects, referred to as "non/occasional-third quartile spenders", were infrequent users of outpatient medical services despite showing an increasing spending pattern over time. Around 33.82% of participants fell into class 3, which were characterized as "heavy-third quartile spenders". This group showed a relatively high probability of spending and maintained a high level of spending throughout the five-year period. All the five time indicators exhibited non-negligible positive relationships with spending. Individuals in this group were characterized as frequent users of outpatient medical services with stably high expenditure trends. The fourth class, comprising an estimated 32.11% of participants, demonstrated a low initial probability of spending but a high baseline upper level of expenditure proportions. The utilization of outpatient medical services was minimal in year 1 but increased rapidly until years 3 and 4, followed by a sharp decrease in year 5. This class is designated as "moderate-third quartile spenders".

As illustrated in Tables 4.6-4.8, concerning the fitted 0.25-quantile Kumaraswamy mixed model, class 1 exhibited a moderate positive correlation among the random intercepts ($\rho_1 = 0.378$), whereas classes 2 and 3 displayed slight negative correlations ($\rho_2 = -0.108$ and $\rho_3 = -0.017$, respectively). On the other hand, class 4 showcased a high positive correlation ($\rho_4 = 0.599$), indicating a strong association between the probability of spending and the 25th percentile of proportional outpatient expenditures. Regarding the estimated median Kumaraswamy mixed model, classes 1 and 3 exhibited moderate positive correlations ($\rho_1 = 0.399$ and $\rho_3 = 0.273$), while class 2 demonstrated a slight negative correlation ($\rho_2 = -0.082$). In contrast, class 4 displayed a relatively high positive correlation ($\rho_4 = 0.539$). In the case of the 0.75-quantile Kumaraswamy mixed model, classes 1 and 2 showed slight positive correlations ($\rho_1 = \rho_2 = 0.141$), whereas class 4 revealed a slight negative correlation

($\rho_4 = -0.151$). Class 3 demonstrated a moderate positive correlation among the random intercepts ($\rho_3 = 0.352$).

4.5 Chapter summary

This chapter introduces a two-part latent class Kumaraswamy quantile mixed regression with Bayesian inference for bounded longitudinal data that exhibit a large spike at zeros. Correlated random effects with class-specific covariance structures are formulated for the binary and the bounded positive components to account for both zero inflation and unobserved heterogeneity. The proposed approach comes with several favourable characteristics. First, under the reparameterization of Kumaraswamy distribution in terms of an assigned quantile and a precision parameter, the model permits the consideration of distinct factors such as time-varying or categorical covariates which can be linked directly to any quantile of interest in the population distribution and allows for individual characteristic attributes which impact the assignment to class membership. Second, the developed method portrays the trajectory of distinct latent class evolutions in the underlying outcome process, which provides valuable insights into the latent cluster structure at various quantiles encompassing the tails and caters to the exploration of skewed longitudinal data with bounded support. Notably, our empirical application convincingly demonstrates that the significance of a covariate may vary across different levels of the response variable within one designated class. In addition, since the posterior distribution is not amenable to analytical solutions, we resort to MCMC estimations which accommodate full posterior inference including HPD regions of parameters and address model uncertainty. Evidenced by the simulation studies, our Bayesian estimators yield desirable results even in the scenario of extreme quantiles (0.75 and 0.95). The application to the RAND HIE on proportional outpatient spending behaviours shows our proposed approach enables the identification of distinct classes of individuals, incorporating the group of spenders who exhibited moderate-to-high probability of spending and the amount spent as well as the group of spenders who showed hesitancy towards such expenses across the lower (0.25-quantile), middle (0.5-quantile) and upper (0.75-quantile) levels of expenditure proportions, respectively. The presented methodology complements the current literature on the analysis of longitudinal data where extreme values are of primary interest.

Chapter 5

Conclusions and future research

The foundation of Bayesian statistics can be traced back to inverse probability (Bayes, 1763) and Bayes' theorem (Laplace, 1814), which have long been established in mathematics but gained significant prominence in applied statistics over the last 50 years. The applications of Bayesian analysis have since flourished across various science-related domains (see Schoot et al., 2021 and references therein for reviews). In sharp contrast to conventional statistics, the Bayesian paradigm treats parameters as random variables, thereby providing a comprehensive quantification of all uncertainties present by means of probability distributions. The principles of Bayesian probability have exhibited remarkable success in the newly emerging natural language processing (NLP) model, ChatGPT. This model combines unsupervised and supervised learning to generate human-like text in response to user input. Bayesian methods provide an elegant approach to managing uncertainty and capturing the inherent structure of the data within ChatGPT. By representing objective functions as probability distributions, the model is empowered to seamlessly integrate and update based on new evidence. Accordingly, ChatGPT can generate responses that are more plausible and coherent, even when confronted with ambiguous or incomplete input.

This thesis has presented several new developments on Bayesian regression models for addressing the challenges related to massive data and extreme longitudinal data exhibiting heavy-tailed characteristics. Clear advantages over existing methods include an adaptive MCMC framework for modelling integer-valued time series with heavy-tailedness, a structure link between Bayesian scale mixtures of normals linear regression and BQR via NIG distribution type of likelihood function, prior and posterior distributions for the calculation of full data posteriors in big data settings and a two-part latent class quantile parametric mixed model with Bayesian inference for skewed longitudinal data with bounded support. The main contributions and future research topics are listed below.

5.1 Main contributions

A Bayesian log-linear Beta–negative binomial integer-valued GARCH process is proposed in Chapter 2. Parameter estimations are performed within adaptive Markov chain Monte Carlo paradigms. The conditions for the posterior distribution of the full model parameter to be proper given some general priors have been derived and presented. In contrast to existing frequentist approaches to modelling discrete time series with heavy-tailedness, the established Bayesian estimators provide a natural way to quantify uncertainty through characterizing the entire posterior distribution and enable model comparison and selection within a unified probabilistic framework.

Chapter 3 contributes to a new approach of Bayesian quantile regression for big data and variable selection. A structure link between Bayesian scale mixtures of normals linear regression and BQR via NIG distribution type of likelihood function, prior distribution and posterior distribution is introduced. The posterior predictive distributions are presented and efficient divide-and-conquer algorithms for BQR and Bayesian LASSO quantile regression are provided.

In Chapter 4, a Bayesian two-part latent class Kumaraswamy quantile mixed model for bounded longitudinal data that exhibit a large spike at zeros is developed. Correlated random effects with class-specific covariance structures are formulated for the binary and the positive components to account for both zero inflation and unobserved heterogeneity. The presented methodology complements the current literature on the analysis of longitudinal data where extreme values are of primary interest.

5.2 Recommendations for future research

In Chapter 2, we considered the application of adaptive MCMC methods for sampling from complex posterior distributions. However, despite achieving convergence, we observed that the MCMC chains may still suffer from high autocorrelation, leading to inefficient exploration of the posterior distribution and increased computational burden. To address the issue of high auto-correlation in MCMC chains, several strategies and algorithms could be potentially considered in our future research. One promising direction is the development of Hamiltonian Monte Carlo (HMC) and its variants, such as the No-U-Turn Sampler (NUTS). HMC employs Hamiltonian dynamics to generate proposals, which can lead to more effective exploration of the posterior distribution compared to traditional random-walk-based MCMC methods (Neal et al., 2011). By simulating trajectories through the parameter space guided by Hamiltonian dynamics, HMC can produce more independent samples with lower autocorrelation. Another avenue of exploration involves the utilization of Sequential Monte Carlo (SMC) methods. SMC, also known as particle filtering, simulates a sequence of weighted particles representing the posterior distribution, with each particle evolving through a series of importance sampling and resampling steps (Doucet et al., 2001). By adaptively updating the particle set, SMC algorithms can effectively explore complex and multimodal posterior distributions, potentially reducing autocorrelation in the MCMC samples.

The work presented in Chapter 2 can be further extended to the threshold model of Tong (1978, 2012), with the intensity process λ_t defined as follows:

$$\lambda_t = \begin{cases} \omega^{(1)} + \phi^{(1)} y_{t-1} + \tau^{(1)} \lambda_{t-1}, & \text{if} \quad y_{t-d} \leqslant c, \\ \omega^{(2)} + \phi^{(2)} y_{t-1} + \tau^{(2)} \lambda_{t-1}, & \text{if} \quad y_{t-d} > c, \end{cases}$$

where y_{t-d} is the threshold variable determining the dynamic switching mechanism of the model, d is a delay lag and c is the threshold value. The threshold model enjoys a piecewise linear path property and is able to capture the dynamic behaviours of time series by a flexible regime-switching framework. The threshold BNB-INGARCH model accommodates both the characterization of heavy-tailedness and the detection of structural changes, which are pivotal to pinpointing since thick tail phenomena and frequent changes in the data-generating mechanism are often encountered owing to instabilities in the real world.

Robust inference is categorized as one of the eight groundbreaking statistical ideas of the past 50 years (Gelman and Vehtari, 2021). Apart from the check loss function tailored to quantile regression given in Chapter 3, other loss functions arising from the M-estimation family, such as the popular Huber loss (Huber, 1964), have received increasing attention. The Huber loss is defined as

$$H_\delta(u) = \begin{cases} u^2/2, & \text{if} \quad |u| \leqslant \delta, \\ \delta|u| - \delta^2/2, & \text{if} \quad |u| > \delta, \end{cases}$$

where $\delta > 0$ is a tuning constant termed as the robustification parameter. The Huber loss function features outlier-robustness of the absolute loss for the least absolute deviation (LAD) regression while maintaining analytical tractability of the squared loss for the least squares. The regression paradigm associated with the Huber loss is referred to as Huber regression. As one of the milestones of robust statistics, Huber regression presents a feasible alternative to quantile regression and leads to the development of various subsequent M-estimators. To date, little attention has been paid to Huber regression in big data analysis. Specifically, Huber robust regression with variable selection methods for distributed massive data represents an area of much research meaning.

We envision future works in Chapter 4 by treating the number of latent classes K as an unknown parameter to estimate. In the endeavour to estimate the number of latent classes, we employed a commonly used model-comparison technique reliant on DIC. Although widely utilized, this method presents certain drawbacks. Notably, it necessitates multiple iterations of the analysis to compare models encompassing differing numbers of classes, resulting in a time-consuming process. Furthermore, the DIC-based approach lacks the capacity to integrate the inherent uncertainty surrounding the determination of the number of classes K, thus potentially overlooking critical nuances within the data structure. Alternative methodologies exist within the realm of Bayesian analysis that offer promising avenues for addressing these limitations. Bayesian nonparametric techniques offer a flexible framework for estimating the latent structure, particularly through methods such as the Dirichlet process (DP) mixture model (Antoniak, 1974). DP mixture models are particularly advantageous in scenarios where the true number of classes is uncertain, offering a probabilistic framework for inferring the underlying class structure from the data (Teh et al., 2006). Other options, such as the Indian Buffet Process (IBP) or Hierarchical Dirichlet Process (HDP), also offer the flexibility to automatically determine the number of latent classes from the data, thus mitigating the need for predefined class specifications (Gershman and Blei, 2012). In addition to these methodologies, other approaches within Bayesian analysis, such as reversible jump Markov chain Monte Carlo (RJMCMC) algorithms which automatically determine the number of latent classes by exploring the model space efficiently, also present viable alternatives for inferring latent class structures while simultaneously accommodating uncertainty in the number of classes (Green, 1995).

Bibliography

Aban, I. B., Meerschaert, M. M., and Panorska, A. K. (2006). Parameter estimation for the truncated Pareto distribution. In: *Journal of the American Statistical Association* 101.473, pp. 270–277.

Ahmad, A. and Francq, C. (2016). Poisson QMLE of count time series models. In: *Journal of Time Series Analysis* 37.3, pp. 291–314.

Albert, J. H. and Chib, S. (1993). Bayesian analysis of binary and polychotomous response data. In: *Journal of the American statistical Association* 88.422, pp. 669–679.

Alhamzawi, R. and Yu, K. (2013). Conjugate priors and variable selection for Bayesian quantile regression. In: *Computational Statistics & Data Analysis* 64, pp. 209–219.

Andrews, D. F. and Mallows, C. L. (1974). Scale Mixtures of Normal Distributions. In: *Journal of the Royal Statistical Society: Series B (Methodological)* 36, pp. 99–102.

Antoniak, C. E. (1974). Mixtures of Dirichlet processes with applications to Bayesian nonparametric problems. In: *The annals of statistics*, pp. 1152–1174.

Aron-Dine, A., Einav, L., and Finkelstein, A. (2013). The RAND health insurance experiment, three decades later. In: *Journal of Economic Perspectives* 27.1, pp. 197–222.

Asparouhov, T. and Muthén, B. (2011). Using Bayesian priors for more flexible latent class analysis. In: *proceedings of the 2011 joint statistical meeting, Miami Beach, FL*. American Statistical Association Alexandria, VA.

Barndorff-Nielsen, O. (1983). On a formula for the distribution of the maximum likelihood estimator. In: *Biometrika* 70.2, pp. 343–365.

Barndorff-Nielsen, O. E. and Shephard, N. (2001). Non-Gaussian Ornstein–Uhlenbeck-based models and some of their uses in financial economics. In: *Journal of the Royal Statistical Society: Series B (Statistical Methodology)* 63, pp. 167–241.

Barreto-Souza, W. (2019). Mixed poisson INAR (1) processes. In: *Statistical Papers* 60.6, pp. 2119–2139.

Bayer, F. M., Cribari-Neto, F., and Santos, J. (2021). Inflated Kumaraswamy regressions with application to water supply and sanitation in Brazil. In: *Statistica Neerlandica* 75.4, pp. 453–481.

Bayes, C. L., Bazán, J. L., and De Castro, M. (2017). A quantile parametric mixed regression model for bounded response variables. In: *Statistics and its interface* 10.3, pp. 483–493.

Bayes, T. (1763). LII. An essay towards solving a problem in the doctrine of chances. By the late Rev. Mr. Bayes, FRS communicated by Mr. Price, in a letter to John Canton, AMFR S. In: *Philosophical transactions of the Royal Society of London* 53, pp. 370–418.

Bernardi, M., Gayraud, G., and Petrella, L. (2015). Bayesian Tail Risk Interdependence Using Quantile Regression. In: *Bayesian Anal.* 10, pp. 553–603.

Bowsher, C. G. and Swain, P. S. (2012). Identifying sources of variation and the flow of information in biochemical networks. In: *Proceedings of the National Academy of Sciences* 109, E1320–E1328.

Briollais, L. and Durrieu, G. (2014). Application of quantile regression to recent genetic and -omic studies. In: *Human genetics* 133, pp. 951–966.

Cappé, O., Moulines, E., Pesquet, J.-C., Petropulu, A. P., and Yang, X. (2002). Long-range dependence and heavy-tail modeling for teletraffic data. In: *IEEE signal processing magazine* 19.3, pp. 14–27.

Celeux, G., Forbes, F., Robert, C., and Titterington, D. (2006). Deviance Information Criteria for Missing Data Models. In: *Bayesian Analysis* 1.4, pp. 651–674.

Chai, H. S. and Bailey, K. R. (2008). Use of log-skew-normal distribution in analysis of continuous data with a discrete component at zero. In: *Statistics in medicine* 27.18, pp. 3643–3655.

Chan, K. and Ledolter, J. (1995). Monte Carlo EM Estimation for Time Series Models Involving Counts. In: *Journal of the American Statistical Association* 90, pp. 242–252.

Chen, C. W. and Khamthong, K. (2020). Bayesian modelling of nonlinear negative binomial integer-valued GARCHX models. In: *Statistical Modelling* 20, pp. 537–561.

Chen, C. W., Khamthong, K., and Lee, S. (2019a). Markov switching integer-valued generalized auto-regressive conditional heteroscedastic models for dengue counts. In: *Journal of the Royal Statistical Society: Series C (Applied Statistics)* 68, pp. 963–983.

Chen, C. W. and Lee, S. (2017). Bayesian causality test for integer-valued time series models with applications to climate and crime data. In: *Journal of the Royal Statistical Society Series C* 66, pp. 797–814.

Chen, C. W. and So, M. K. (2006). On a threshold heteroscedastic model. In: *International Journal of Forecasting* 22, pp. 73–89.

Chen, X., Liu, W., Mao, X., and Yang, Z. (2020). Distributed high-dimensional regression under a quantile loss function. In: *Journal of Machine Learning Research* 21.

Chen, X., Liu, W., and Zhang, Y. (2019b). Quantile regression under memory constraint. In: *Ann. Statist.* 47, pp. 3244–3273.

Chib, S. and Winkelmann, R. (2001). Markov Chain Monte Carlo Analysis of Correlated Count Data. In: *Journal of Business & Economic Statistics* 19, pp. 428–435.

Christou, V. and Fokianos, K. (2014). Quasi-likelihood inference for negative binomial time series models. In: *Journal of Time Series Analysis* 35, pp. 55–78.

Cole, T. and Green, P. (1992). Smoothing reference centile curves: the LMS method and penalized likelihood. In: *Statistics in medicine* 11, pp. 1305–1319.

Contreras, J., Espinola, R., Nogales, F. J., and Conejo, A. J. (2003). ARIMA models to predict next-day electricity prices. In: *IEEE Transactions on Power Systems* 18, pp. 1014–1020.

Cook, D. O., Kieschnick, R., and McCullough, B. D. (2008). Regression analysis of proportions in finance with self selection. In: *Journal of empirical finance* 15.5, pp. 860–867.

Cox, D. R., Gudmundsson, G., Lindgren, G., Bondesson, L., Harsaae, E., Laake, P., Juselius, K., and Lauritzen, S. L. (1981). Statistical Analysis of Time Series: Some Recent Developments [with Discussion and Reply]. In: *Scandinavian Journal of Statistics* 8, pp. 93–115.

Cragg, J. G. (1971). Some statistical models for limited dependent variables with application to the demand for durable goods. In: *Econometrica: Journal of the Econometric Society*, pp. 829–844.

Cribari-Neto, F. and Santos, J. (2019). Inflated Kumaraswamy distributions. In: *Anais da Academia Brasileira de Ciências* 91.

Dai, H., Bao, Y., and Bao, M. (2013). Maximum likelihood estimate for the dispersion parameter of the negative binomial distribution. In: *Statistics & Probability Letters* 83.1, pp. 21–27.

Davis, R. A., Dunsmuir, W. T., and Streett, S. B. (2003). Observation-driven models for Poisson counts. In: *Biometrika* 90, pp. 777–790.

— (2005). Maximum Likelihood Estimation for an Observation Driven Model for Poisson Counts. In: *Methodol Comput Appl Probab* 7, pp. 149–159.

Davis, R. A., Holan, S. H., Lund, R., and Ravishanker, N. (2016). Handbook of discrete-valued time series. CRC Press.

Deb, P. and Trivedi, P. K. (2002). The structure of demand for health care: latent class versus two-part models. In: *Journal of health economics* 21.4, pp. 601–625.

Douc, R., Doukhan, P., and Moulines, E. (2013). Ergodicity of observation-driven time series models and consistency of the maximum likelihood estimator. In: *Stochastic Processes and their Applications* 123.7. A Special Issue on the Occasion of the 2013 International Year of Statistics, pp. 2620–2647.

Doucet, A., De Freitas, N., and Gordon, N. (2001). An introduction to sequential Monte Carlo methods. In: *Sequential Monte Carlo methods in practice*, pp. 3–14.

Doukhan, P., Fokianos, K., and Tjøstheim, D. (2012). On weak dependence conditions for Poisson autoregressions. In: *Statistics & Probability Letters* 82, pp. 942–948.

Drescher, D. (2005). Alternative distributions for observation driven count series models. Economics Working Paper 2005-11.

Duan, N., Manning, W. G., Morris, C. N., and Newhouse, J. P. (1983). A comparison of alternative models for the demand for medical care. In: *Journal of business & economic statistics* 1.2, pp. 115–126.

Dunsmuir, W. T. and Scott, D. J. (2015). The glarma Package for Observation-Driven Time Series Regression of Counts. In: *Journal of Statistical Software* 67, pp. 1–36.

Durbin, J. and Koopman, S. J. (2000). Time series analysis of non-Gaussian observations based on state space models from both classical and Bayesian perspectives. In: *Journal of the Royal Statistical Society: Series B (Statistical Methodology)* 62, pp. 3–56.

Elliott, M. R., Gallo, J. J., Ten Have, T. R., Bogner, H. R., and Katz, I. R. (2005). Using a Bayesian latent growth curve model to identify trajectories of positive affect and negative events following myocardial infarction. In: *Biostatistics* 6.1, pp. 119–143.

Fahrmeir, L., Tutz, G., Hennevogl, W., and Salem, E. (2013). Multivariate Statistical Modelling Based on Generalized Linear Models. Springer Series in Statistics. Springer New York.

Fan, J., Fan, Y., and Barut, E. (2014). Adaptive robust variable selection. In: *Annals of statistics* 42.1, p. 324.

Ferland, R., Latour, A., and Oraichi, D. (2006). Integer-Valued GARCH Process. In: *Journal of Time Series Analysis* 27, pp. 923–942.

Ferrari, S. and Cribari-Neto, F. (2004). Beta regression for modelling rates and proportions. In: *Journal of applied statistics* 31.7, pp. 799–815.

Figueroa-Zúñiga, J., Niklitschek-Soto, S., Leiva, V., and Liu, S. (2022). Modeling heavy-tailed bounded data by the trapezoidal beta distribution with applications. In: *REVSTAT-Statistical Journal* 20.3, pp. 387–404.

Fokianos, K. (2012). 12 - Count Time Series Models. In: *Time Series Analysis: Methods and Applications*. Vol. 30. Handbook of Statistics. Elsevier, pp. 315–347.

Fokianos, K. and Tjøstheim, D. (2011). Log-linear Poisson autoregression. In: *Journal of Multivariate Analysis* 102.3, pp. 563–578.

Frankenberg, C., Thorpe, A. K., Thompson, D. R., Hulley, G., Kort, E. A., Vance, N., Borchardt, J., Krings, T., Gerilowski, K., Sweeney, C., et al. (2016). Airborne methane remote measurements reveal heavy-tail flux distribution in Four Corners region. In: *Proceedings of the national academy of sciences* 113.35, pp. 9734–9739.

Frühwirth-Schnatter, S. (2001). Markov chain Monte Carlo Estimation of Classical and Dynamic Switching and Mixture Models. In: *Journal of the American Statistical Association* 96.453, pp. 194–209.

— (2006). Finite mixture and Markov switching models. Springer.

Gelman, A., Carlin, J., Stern, H., Dunson, D., Vehtari, A., and Rubin, D. (2013). Bayesian Data Analysis, Third Edition. Chapman & Hall/CRC Texts in Statistical Science. Taylor & Francis.

Gelman, A., Hwang, J., and Vehtari, A. (2014). Understanding predictive information criteria for Bayesian models. In: *Statistics and computing* 24.6, pp. 997–1016.

Gelman, A., Roberts, G. O., Gilks, W. R., et al. (1996). Efficient Metropolis jumping rules. In: *Bayesian statistics* 5.599–608, p. 42.

Gelman, A. and Vehtari, A. (2021). What are the most important statistical ideas of the past 50 years? In: *Journal of the American Statistical Association* 116.536, pp. 2087–2097.

Gershman, S. J. and Blei, D. M. (2012). A tutorial on Bayesian nonparametric models. In: *Journal of Mathematical Psychology* 56.1, pp. 1–12.

Gonçalves, K. C., Migon, H. S., and Bastos, L. S. (2020). Dynamic quantile linear models: A bayesian approach. In: *Bayesian Analysis* 15.2, pp. 335–362.

Gorgi, P. (2020). Beta–negative binomial auto-regressions for modelling integer-valued time series with extreme observations. In: *Journal of the Royal Statistical Society: Series B (Statistical Methodology)* 82, pp. 1325–1347.

Green, P. J. (1995). Reversible jump Markov chain Monte Carlo computation and Bayesian model determination. In: *Biometrika* 82.4, pp. 711–732.

Gu, Y., Fan, J., Kong, L., Ma, S., and Zou, H. (2018). ADMM for high-dimensional sparse penalized quantile regression. In: *Technometrics* 60, pp. 319–331.

Gupta, M., Qu, P., and Ibrahim, J. G. (2007). A temporal hidden Markov regression model for the analysis of gene regulatory networks. In: *Biostatistics* 8, pp. 805–820.

Haario, H., Saksman, E., and Tamminen, J. (1999). Adaptive proposal distribution for random walk Metropolis algorithm. In: *Computational statistics* 14, pp. 375–395.

— (2001). An adaptive Metropolis algorithm. In: *Bernoulli*, pp. 223–242.

Hahn, E. D. (2008). Mixture densities for project management activity times: A robust approach to PERT. In: *European Journal of operational research* 188.2, pp. 450–459.

Han, J., Slate, E. H., and Peña, E. A. (2007). Parametric latent class joint model for a longitudinal biomarker and recurrent events. In: *Statistics in medicine* 26.29, pp. 5285–5302.

Harvey, A. and Luati, A. (2014). Filtering With Heavy Tails. In: *Journal of the American Statistical Association* 109, pp. 1112–1122.

Hay, J. L. and Pettitt, A. N. (2001). Bayesian analysis of a time series of counts with covariates: an application to the control of an infectious disease. In: *Biostatistics* 2, pp. 433–444.

Heckman, J. J. (1976). The common structure of statistical models of truncation, sample selection and limited dependent variables and a simple estimator for such models. In: *Annals of economic and social measurement, volume 5, number 4*. NBER, pp. 475–492.

Holla, M. (1967). On a Poisson-inverse Gaussian distribution. In: *Metrika* 11.1, pp. 115–121.

Huang, Q., Zhang, H., Chen, J., and He, M. (2017). Quantile regression models and their applications: A review. In: *Journal of Biometrics & Biostatistics* 8.3, pp. 1–6.

Huber, P. J. (1964). Robust Estimation of a Location Parameter. In: *The Annals of Mathematical Statistics* 35.1, pp. 73–101.

— (2011). Robust statistics. In: *International encyclopedia of statistical science*. Springer, pp. 1248–1251.

Johnson, N. L., Kemp, A. W., and Kotz, S. (2005). Univariate discrete distributions. Vol. 444. John Wiley & Sons.

Jones, M. (2009). Kumaraswamy's distribution: A beta-type distribution with some tractability advantages. In: *Statistical methodology* 6.1, pp. 70–81.

Jung, R. C., Kukuk, M., and Liesenfeld, R. (2006). Time series of count data: modeling, estimation and diagnostics. In: *Computational Statistics & Data Analysis* 51, pp. 2350–2364.

Kedem, B. and Fokianos, K. (2005). Regression Models for Time Series Analysis. Wiley Series in Probability and Statistics. Wiley.

Khashei, M., Bijari, M., and Ardali, G. A. R. (2009). Improvement of Auto-Regressive Integrated Moving Average models using Fuzzy logic and Artificial Neural Networks (ANNs). In: *Neurocomputing* 72, pp. 956–967.

Kim, S., Caporaso, N. E., Gu, F., Klerman, E. B., and Albert, P. S. (2023). Uncovering circadian rhythms in metabolic longitudinal data: A Bayesian latent class modeling approach. In: *Statistics in Medicine* 42.18, pp. 3302–3315.

Koenker, R. and Hallock, K. (2001). Quantile regression: An introduction. In: *Journal of Economic Perspectives* 15, pp. 43–56.

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

— (1978b). Regression quantiles. In: *Econometrica*, pp. 33–50.

Kozumi, H. and Kobayashi, G. (2011a). Gibbs sampling methods for Bayesian quantile regression. In: *Journal of statistical computation and simulation* 81.11, pp. 1565–1578.

— (2011b). Gibbs sampling methods for Bayesian quantile regression. In: *J. Stat. Comput. Simul.* 81.11, pp. 1565–1578.

Kuk, A. Y. and Cheng, Y. W. (1997). The monte carlo newton-raphson algorithm. In: *Journal of Statistical Computation and Simulation* 59, pp. 233–250.

Kumaraswamy, P. (1976). Sinepower probability density function. In: *Journal of Hydrology* 31.1-2, pp. 181–184.

Kumaraswamy, P. (1980). A generalized probability density function for double-bounded random processes. In: *Journal of hydrology* 46.1-2, pp. 79–88.

Kysely, J. and Picek, J. (2007). Regional growth curves and improved design value estimates of extreme precipitation events in the Czech Republic. In: *Climate research* 33.3, pp. 243–255.

Lambert, D. (1992). Zero-inflated Poisson regression, with an application to defects in manufacturing. In: *Technometrics* 34.1, pp. 1–14.

Laplace, P.-S. (1814). Essai philosophique sur les probabilités (A philosophical essay on probabilities). In: *Paris, France: Veuve Courcier.*

Lee, C.-M. and Ko, C.-N. (2011). Short-term load forecasting using lifting scheme and ARIMA models. In: *Expert Systems with Applications* 38, pp. 5902–5911.

Lee, K. E., Sha, N., Dougherty, E. R., Vannucci, M., and Mallick, B. K. (2003). Gene selection: a Bayesian variable selection approach. In: *Bioinformatics* 19, pp. 90–97.

Leiby, B. E., Sammel, M. D., Ten Have, T. R., and Lynch, K. G. (2009). Identification of multivariate responders and non-responders by using Bayesian growth curve latent class models. In: *Journal of the Royal Statistical Society Series C: Applied Statistics* 58.4, pp. 505–524.

Lenk, P. J. and DeSarbo, W. S. (2000). Bayesian inference for finite mixtures of generalized linear models with random effects. In: *Psychometrika* 65.1, pp. 93–119.

Leroux, B. G. and Puterman, M. L. (1992). Maximum-Penalized-Likelihood Estimation for Independent and Markov- Dependent Mixture Models. In: *Biometrics* 48, pp. 545–558.

Lévy, P. (1925). Calcul des probabilités. Gauthier-Villars.

Li, Y. and Zhu, J. (2008). $L1$-Norm Quantile Regression. In: *Journal of Computational and Graphical Statistics* 17, pp. 163–185.

Liboschik, T., Fokianos, K., and Fried, R. (2017). tscount: An R package for analysis of count time series following generalized linear models. In: *Journal of Statistical Software* 82, pp. 1–51.

Lin, H., Turnbull, B. W., McCulloch, C. E., and Slate, E. H. (2002). Latent class models for joint analysis of longitudinal biomarker and event process data: application to longitudinal prostate-specific antigen readings and prostate cancer. In: *Journal of the American Statistical Association* 97.457, pp. 53–65.

Liu, F. and Eugenio, E. C. (2018). A review and comparison of Bayesian and likelihood-based inferences in beta regression and zero-or-one-inflated beta regression. In: *Statistical methods in medical research* 27.4, pp. 1024–1044.

Liu, K. (1993). A new class of blased estimate in linear regression. In: *Communications in Statistics-Theory and Methods* 22.2, pp. 393–402.

Liu, L., Strawderman, R. L., Cowen, M. E., and Shih, Y.-C. T. (2010). A flexible two-part random effects model for correlated medical costs. In: *Journal of health economics* 29.1, pp. 110–123.

Lord, D. and Geedipally, S. R. (2018). Safety prediction with datasets characterised with excess zero responses and long tails. In: *Safe Mobility: Challenges, Methodology and Solutions*. Vol. 11. Emerald Publishing Limited, pp. 297–323.

Lu, S. and Molz, F. J. (2001). How well are hydraulic conductivity variations approximated by additive stable processes? In: *Advances in Environmental Research* 5.1, pp. 39–45.

Manning, W. G., Newhouse, J. P., Duan, N., Keeler, E. B., and Leibowitz, A. (1987). Health insurance and the demand for medical care: evidence from a randomized experiment. In: *The American economic review*, pp. 251–277.

Maronna, R. A., Martin, R. D., Yohai, V. J., and Salibián-Barrera, M. (2019). Robust statistics: theory and methods (with R). John Wiley & Sons.

McCabe, B. P. and Martin, G. M. (2005). Bayesian predictions of low count time series. In: *International Journal of Forecasting* 21, pp. 315–330.

Mitnik, P. A. and Baek, S. (2013). The Kumaraswamy distribution: median-dispersion re-parameterizations for regression modeling and simulation-based estimation. In: *Statistical Papers* 54.1, pp. 177–192.

Mullahy, J. (1986). Specification and testing of some modified count data models. In: *Journal of econometrics* 33.3, pp. 341–365.

Nadarajah, S. (2008). On the distribution of Kumaraswamy. In: *Journal of Hydrology* 348.3, pp. 568–569.

Nagahara, Y. (1999). The PDF and CF of Pearson type IV distributions and the ML estimation of the parameters. In: *Statistics & probability letters* 43.3, pp. 251–264.

Nariswari, R. and Pudjihastuti, H. (2019). Bayesian Forecasting for Time Series of Count Data. In: *Procedia Computer Science* 157, pp. 427–435.

Neal, R. M. et al. (2011). MCMC using Hamiltonian dynamics. In: *Handbook of markov chain monte carlo* 2.11, p. 2.

Neelon, B., O'Malley, A. J., and Normand, S.-L. T. (2011a). A Bayesian two-part latent class model for longitudinal medical expenditure data: assessing the impact of mental health and substance abuse parity. In: *Biometrics* 67.1, pp. 280–289.

Neelon, B., Swamy, G. K., Burgette, L. F., and Miranda, M. L. (2011b). A Bayesian growth mixture model to examine maternal hypertension and birth outcomes. In: *Statistics in medicine* 30.22, pp. 2721–2735.

Neelon, B., Zhu, L., and Neelon, S. E. B. (2015). Bayesian two-part spatial models for semicontinuous data with application to emergency department expenditures. In: *Biostatistics* 16.3, pp. 465–479.

Neumann, M. H. (2011). Absolute regularity and ergodicity of Poisson count processes. In: *Bernoulli* 17, pp. 1268–1284.

Nobile, A. (2004). On the Posterior Distribution of the Number of Components in a Finite Mixture. In: *Annals of Statistics*, pp. 2044–2073.

Ospina, R. and Ferrari, S. L. (2010). Inflated beta distributions. In: *Statistical papers* 51.1, pp. 111–126.

Park, T. and Casella, G. (2008). The Bayesian Lasso. In: *Journal of the American Statistical Association* 103, pp. 681–686.

Petrella, L. and Raponi, V. (2019). Joint estimation of conditional quantiles in multivariate linear regression models with an application to financial distress. In: *Journal of Multivariate Analysis* 173.C, pp. 70–84.

Pirmohammadi, S. and Bidram, H. (2022). On the Liu estimator in the beta and Kumaraswamy regression models: A comparative study. In: *Communications in Statistics-Theory and Methods* 51.24, pp. 8553–8578.

Proust-Lima, C., Séne, M., Taylor, J. M., and Jacqmin-Gadda, H. (2014). Joint latent class models for longitudinal and time-to-event data: a review. In: *Statistical methods in medical research* 23.1, pp. 74–90.

Qian, L., Li, Q., and Zhu, F. (2020). Modelling heavy-tailedness in count time series. In: *Applied Mathematical Modelling* 82, pp. 766–784.

R Core Team (2021). R: A Language and Environment for Statistical Computing. R Foundation for Statistical Computing. Vienna, Austria. URL: https://www.R-project.org/.

Reed, C. and Yu, K. (2009). A partially collapsed Gibbs sampler for Bayesian quantile regression. Tech. rep. Department of Mathematical Sciences, Brunel University.

Resnick, S. I. (1997). Heavy tail modeling and teletraffic data: special invited paper. In: *The Annals of Statistics* 25.5, pp. 1805–1869.

Richardson, S. and Green, P. J. (1997). On Bayesian Analysis of Mixtures with an Unknown Number of Components. In: *Journal of the Royal Statistical Society. Series B (Methodological)* 59.4, pp. 731–792.

Rodrigues, T. and Fan, Y. (2017). Regression adjustment for noncrossing Bayesian quantile regression. In: *Journal of Computational and Graphical Statistics* 26, pp. 275–284.

Roth, M. (2012). On the Multivariate t Distribution. Tech. rep. 3059. Linköping University, The Institute of Technology, p. 22.

Rousseeuw, P. J. and Hubert, M. (2011). Robust statistics for outlier detection. In: *Wiley interdisciplinary reviews: Data mining and knowledge discovery* 1.1, pp. 73–79.

Schifano, E. D., Wu, J., Wang, C., Yan, J., and Chen, M.-H. (2016). Online Updating of Statistical Inference in the Big Data Setting. In: *Technometrics* 58.3, pp. 393–403.

Schoot, R. van de, Depaoli, S., King, R., Kramer, B., Märtens, K., Tadesse, M. G., Vannucci, M., Gelman, A., Veen, D., Willemsen, J., et al. (2021). Bayesian statistics and modelling. In: *Nature Reviews Methods Primers* 1.1, p. 1.

Scotto, M. G., Weiß, C. H., and Gouveia, S. (2015). Thinning-based models in the analysis of integer-valued time series: a review. In: *Statistical Modelling* 15, pp. 590–618.

Sherwood, B. and Wang, L. (2016). Partially linear additive quantile regression in ultra-high dimension. In: *The Annals of Statistics* 44.1, pp. 288–317.

Silva, R. B. and Barreto-Souza, W. (2019). Flexible and robust mixed Poisson INGARCH models. In: *Journal of Time Series Analysis* 40.5, pp. 788–814.

Silveira de Andrade B, Andrade, M. G., and Ehlers, R. S. (2015). Bayesian GARMA models for count data. In: *Communications in Statistics: Case Studies, Data Analysis and Applications* 1, pp. 192–205.

Singh, A., Van Dorp, J. R., and Mazzuchi, T. A. (2007). A novel asymmetric distribution with power tails. In: *Communications in Statistics—Theory and Methods* 36.2, pp. 235–252.

Smith, M. and Kohn, R. (1996). Nonparametric regression using Bayesian variable selection. In: *Journal of Econometrics* 75, pp. 317–343.

Spiegelhalter, D. J., Best, N. G., Carlin, B. P., and Van Der Linde, A. (2002). Bayesian measures of model complexity and fit. In: *Journal of the royal statistical society: Series b (statistical methodology)* 64.4, pp. 583–639.

Stephens, M. (2000). Dealing with label switching in mixture models. In: *Journal of the Royal Statistical Society: Series B (Statistical Methodology)* 62.4, pp. 795–809.

Su, L., Tom, B. D., and Farewell, V. T. (2009). Bias in 2-part mixed models for longitudinal semicontinuous data. In: *Biostatistics* 10.2, pp. 374–389.

Taylor, J. W. (2017). Probabilistic forecasting of wind power ramp events using autoregressive logit models. In: *European Journal of Operational Research* 259.2, pp. 703–712.

Teh, Y., Jordan, M., Beal, M., and Blei, D. (2006). In: *Journal of the American Statistical Association* 101.476, pp. 1566–1581.

Tibshirani, R. (1996). Regression Shrinkage and Selection Via the Lasso. In: *Journal of the Royal Statistical Society: Series B (Methodological)* 58, pp. 267–288.

Tjøstheim, D. (2016). Count time series with observation-driven autoregressive parameter dynamics. In: *Handbook of Discrete-Valued Time Series*. Chapman & Hall/CRC Handbooks of Modern Statistical Methods. CRC Press, pp. 77–100.

Tong, H. (1978). On a threshold model in pattern recognition and signal processing, ed. In: *CH Chen, Amsterdam: Sijhoff & Noordhoff.*

— (2012). Threshold models in non-linear time series analysis. Vol. 21. Springer Science & Business Media.

Truong, B.-C., Chen, C. W., and Sriboonchitta, S. (2017). Hysteretic Poisson INGARCH model for integer-valued time series. In: *Statistical Modelling* 17, pp. 401–422.

Wang, C., Liu, H., Yao, J.-F., Davis, R. A., and Li, W. K. (2014). Self-excited threshold Poisson autoregression. In: *Journal of the American Statistical Association* 109.506, pp. 777–787.

Wang, C., Chen, M.-H., Schifano, E., Wu, J., and Yan, J. (2016a). Statistical methods and computing for big data. In: *Stat. Its Interface* 9.4, pp. 399–414.

Wang, H., Li, G., and Tsai, C.-L. (2007). Regression coefficient and autoregressive order shrinkage and selection via the lasso. In: *Journal of the Royal Statistical Society Series B: Statistical Methodology* 69.1, pp. 63–78.

Wang, Y., Feng, X.-N., and Song, X.-Y. (2016b). Bayesian quantile structural equation models. In: *Structural Equation Modeling: A Multidisciplinary Journal* 23, pp. 246–258.

Wang, Z. (2011). One mixed negative binomial distribution with application. In: *Journal of Statistical Planning and Inference* 141, pp. 1153–1160.

Weiß, C. H. (2018). An introduction to discrete-valued time series. John Wiley & Sons.

West, M. and Harrison, J. (2006). Bayesian forecasting and dynamic models. Springer Science & Business Media.

White, H. (1980). A Heteroskedasticity-Consistent Covariance Matrix Estimator and a Direct Test for Heteroskedasticity. In: *Econometrica* 48, pp. 817–838.

Willmot, G. E. (1987). The Poisson-inverse Gaussian distribution as an alternative to the negative binomial. In: *Scandinavian Actuarial Journal* 1987.3-4, pp. 113–127.

Winkelmann, R. (2008). Econometric Analysis of Count Data. Springer Berlin Heidelberg.

Wu, Y. and Yin, G. (2015). Conditional quantile screening in ultrahigh-dimensional heterogeneous data. In: *Biometrika* 102, pp. 65–76.

Xiong, L. and Zhu, F. (2019). Robust quasi-likelihood estimation for the negative binomial integer-valued GARCH(1,1) model with an application to transaction counts. In: *Journal of Statistical Planning and Inference* 203, pp. 178–198.

Yang, S. and Puggioni, G. (2021). Bayesian zero-inflated growth mixture models with application to health risk behavior data. In: *Statistics and Its Interface* 14.2, pp. 151–163.

Yu, K. and Stander, J. (2007). Bayesian analysis of a Tobit quantile regression model. In: *Journal of Econometrics* 137, pp. 260–276.

Yu, K., Lu, Z., and Stander, J. (2003). Quantile Regression: Applications and Current Research Areas. In: *Journal of the Royal Statistical Society. Series D (The Statistician)* 52, pp. 331–350.

Yu, K. and Moyeed, R. A. (2001). Bayesian quantile regression. In: *Statistics & Probability Letters* 54.4, pp. 437–447.

Yu, L., Lin, N., and Wang, L. (2017). A parallel algorithm for large-scale nonconvex penalized quantile regression. In: *Journal of Computational and Graphical Statistics* 26, pp. 935–939.

Zabell, S. L. (2008). On student's 1908 article "the probable error of a mean". In: *Journal of the American Statistical Association* 103.481, pp. 1–7.

Zeger, S. L. (1988). A regression model for time series of counts. In: *Biometrika* 75, pp. 621–629.

Zeger, S. L. and Qaqish, B. (1988). Markov Regression Models for Time Series: A Quasi-Likelihood Approach. In: *Biometrics* 44, pp. 1019–1031.

Zellner, A. (1986). On assessing prior distributions and Bayesian regression analysis with g-prior distributions. In: *In Bayesian Inference and Decision techniques. Stud. Bayesian Econometrics Statist* 6, pp. 233–243.

Zhu, F. (2011). A negative binomial integer-valued GARCH model. In: *Journal of Time Series Analysis* 32, pp. 54–67.

Zou, H. and Yuan, M. (2008). Composite quantile regression and the oracle model selection theory. In: *The Annals of Statistics* 36.3, pp. 1108–1126.