Bayesian mode regression

Journal:	*Scandinavian Journal of Statistics*
Manuscript ID:	SJS-14-188
Manuscript Type:	Paper
Date Submitted by the Author:	14-Aug-2014
Complete List of Authors:	Yu, Keming; Brunel University, Mathematics

Bayesian Mode Regression

Keming Yu, *Brunel University, London, UB8 3PH, UK*
Katerina Aristodemou, *Brunel University, London, UB8 3PH, UK*
Zudi Lu, *University of Southampton, Southampton, SO17 1BJ, UK*

Abstract

Like mean, quantile and variance, mode is also an important measure of central tendency of a distribution. Many practical questions, particularly in the analysis of big data, such as "Which element (gene or file or signal) is the most typical one among all elements in a network?" are directly related to mode. Mode regression, which provides a convenient summary of how the regressors affect the conditional mode, is totally different from other models based on conditional mean or conditional quantile or conditional variance. Some inference methods for mode regression exist but none of them is from the Bayesian perspective. This paper introduces Bayesian mode regression by exploring three different approaches, including their theoretic properties. The proposed approacher are illustrated using simulated datasets and a real data set.

Keywords: Bayesian inference; Empirical likelihood; Mode regression

1. INTRODUCTION

Mode, the most likely value of a distribution, has wide applications in biology, astronomy, economics and finance. In these fields, it is not uncommon to encounter data distributions that are skewed or contain outliers. In those cases, the arithmetic mean may not be an appropriate statistic to represent the center of location of the data. Alternative statistics with less bias are the median and the mode. The mean or the median of two densities may be identical, while the shapes of the two densities can be quite different. The mode preserves some of the important features, such as wiggles, of the underlying distribution function, whereas the mean and the median tend to average out the data.

The mode has been used in modern science to identify the most frequent or the most typical element in certain network systems (Hedges and Shah (2003), Heckman et al. (2001), Kumar and Hedges (1998), Markov et al. (1997)). Mode estimation has attracted significant attention in the statistics literature for decades by various authors [Yasukawa (1926), Parzen (1962), Grenander (1965), Eddy (1980), Bickel and Fan (1996), Birgé (1997), Berlinet et al. (1998) and Meyer (2001) among others]. Moreover, identifying the typical value or pattern could be one of the most efficient statistical approaches for the analysis of big data.

However, mode estimation is more difficult than estimating the mean or the median. The mode estimator is often defined as the maximum of the estimated distribution density, typically under nonparametric kernel estimation. Conditional mode estimation is typically carried out by conditional density estimation via different nonparametric methods [see for example Gasser et al. (1998), Hall and Huang (2001) and Hall et al. (2001), Brunner (1992), Ho (2006), Dunson et al. (2007)].

However, these nonparametric conditional density-based mode regression models do not provide a direct estimate of the conditional mode. The problem with these methods is twofold: the estimation of the conditional density may suffer from the well-known "curse of dimensionality" and, it is hard to describe and interpret the estimated conditional mode in terms of predictors or covariates.

Direct inference for mode regression was explored by Lee first in 1989, Lee (1989), and then in 1993, Lee (1993). However, it has not been well-applied due to lack of proper inference tools. Recently, Kemp and Santos Silva (2012) relaxed Lee's restriction on truncated dependent variables and employed alternative kernel estimation. However, their regression coefficient estimator has slow convergence rate, involves bandwidth selection and provides only approximate Normal confidence intervals. Furthermore, Yao and Li (2013) proposed an Expectation-Maximisation algorithm in order to estimate the regression coefficients of the modal linear regression. These methods involve either semiparametric or nonparametric estimation methods. A direct Bayesian method for mode regression is not available even though there is a clear practical motivation from this perspective.

In conventional regression models, the method of least squares is usually applied to investigate the effect of the predictor variables on the conditional mean of the response variable. However, in the presence of outliers, the mean is pulled in the direction of the tail, making mean regression a less representative method of analysis. Mode regression, on the other hand, is robust to the presence

of outliers. Quantile regression is an alternative approach to estimate models with skewed data, as it can provide a complete picture of the conditional distribution of the response variable given the covariates. However, it cannot reveal any information about the typical value (mode).

Take the analysis of the adult Body Mass Index (BMI) used in this paper as an example. BMI, defined by $BMI = \frac{weight(kg)}{height^2(cm)}$, is a measure of the relative weight and is used in a wide variety of contexts as a simple method to assess how much an individual's body weight deviates from what is normal or desirable for a person of his or her height. Such analysis is important as it is well-known that obesity has overtaken smoking as the biggest threat to people's health, in particular for middle-aged and old adults.

The dataset used in this paper to demonstrate mode regression is taken from the Health Survey for England (HSE) 2011 teaching dataset. The Health Survey for England is a series of annual surveys about the health of people living in England, commissioned by the Department of Health. The sample contains observations for 4,138 individuals (1,814 males and 2,324 females) with two thirds being older than 40 years old. A BMI of $27kg/m^2$ for middle-aged and old adults can be classified as the cut-off point of unhealthy weight. An interesting question is how some covariates, such as units of alcohol and portions of fruit/vegetables consumed keep one's BMI in the healthy range. It would be safe to assume that the BMI for the majority of people in the data example falls in the desirable BMI range. Indeed, the typical BMI for the whole sample as well as separately for men or women are below $27kg/m^2$, but the corresponding mean BMI and median BMI were near or greater than $27kg/m^2$. Therefore, employing mode regression is preferable than mean and quantile regression for answering this scientific question.

In this paper we introduce a fully Bayesian framework for direct mode regression inference by using three approaches: a parametric Bayesian method, a nonparametric Bayesian method and an empirical likelihood based Bayesian method. The remainder of the paper is organized as follows. Section 2 introduces the three approaches, describes the theoretical and computational framework of these methods and gives their mathematical justification. In Section 3 we illustrate the proposed methods through two simulated case-studies and a real example. We conclude with a short discussion in Section 4.

2. BAYESIAN MODE REGRESSION

2.1. Mode Estimation

Consider an arbitrary random variable Z, with distribution function $F_Z(z)$ and density function $f_Z(z)$. Let $K(Z;\cdot)$ be the *step-loss function* (Manski (1991)) such as,

$$K(Z;\mu) = I\left[\frac{|Z-\mu|}{\sigma} > 1\right], \tag{2.1}$$

with $\sigma > 0$ and $I[A]$ being the indicator function of event A. If $f_Z(z)$ is symmetric around μ or if μ is the middle value of the interval of length 2σ that captures the most probability under $F_Z(z)$, then

$$\hat{\mu} = \text{argmin}_\mu E\{K(Z;\mu)\}$$

is the mode of Z.

Therefore, given a sample $\{Z_1, \ldots Z_n\}$ from Z, let $\hat{\mu}$ be the estimator of the mode of Z, then,

$$\hat{\mu} = \text{argmin}_\mu \sum_{i=1}^{n} I\left[|Z_i - \mu| > \sigma\right].$$

Generally, if we define a uniform density function $f_\sigma(u)$ over the interval $(\mu-\sigma,\ \mu+\sigma)$ as

$$f_\sigma(u;\mu) = \frac{1}{2\sigma} I(|u-\mu| \le \sigma), \tag{2.2}$$

where $-\infty < \mu < \infty$ and $\sigma > 0$ are the location parameter and scale parameter respectively, then clearly the mode of Z can be estimated by

$$\hat{\mu} = \text{argmax}_\mu \prod_{i=1}^{n} f_\sigma(Z_i;\mu). \tag{2.3}$$

2.2. Parametric Bayesian Mode Regression

Let $mode(y|\boldsymbol{x}) = \boldsymbol{x}'\boldsymbol{\beta}$ be the conditional mode of Y given $X = x$. A standard regression model to formulate the mode regression could be as:

$$y = \boldsymbol{x}'\boldsymbol{\beta} + \epsilon, \tag{2.4}$$

with $mode(\epsilon|\boldsymbol{x}) = 0$ of model error ϵ.

Lee (1989,1993) showed that, given a sample $\{(\boldsymbol{x}_1, y_1), \ldots (\boldsymbol{x}_n, y_n)\}$ from $(\boldsymbol{x}, y)$, the classical mode regression estimator, $\hat{\boldsymbol{\beta}}$ is given by:

$$\hat{\boldsymbol{\beta}} = \text{argmax}_{\boldsymbol{\beta}} \frac{1}{n} \sum_{i=1}^{n} I[|y_i - \boldsymbol{x}_i'\boldsymbol{\beta}| \leq \sigma]. \tag{2.5}$$

Therefore, using equation (2.3), $\hat{\boldsymbol{\beta}}$ can be regarded as the maximum likelihood estimator of the "working" likelihood function

$$L(y|\boldsymbol{\beta}, \sigma) = \prod_{i=1}^{n} f_\sigma(y_i; x_i'\boldsymbol{\beta}). \tag{2.6}$$

Now a natural joint posterior distribution of the unknown model parameters, $\boldsymbol{\beta}$ and σ under a Bayesian framework is given by

$$\pi(\boldsymbol{\beta}, \sigma|y) \propto L(y|\boldsymbol{\beta}, \sigma)\, \pi(\boldsymbol{\beta}, \sigma), \tag{2.7}$$

where $\pi(\boldsymbol{\beta}, \sigma)$ is the joint prior distribution of $\boldsymbol{\beta}$ and σ.

The Bayesian mode regression estimates, denoted as $\hat{\boldsymbol{\beta}}_B$ can be obtained using the marginal posterior distribution of $\boldsymbol{\beta}$, given by

$$\pi(\boldsymbol{\beta}|y) = \int \pi(\boldsymbol{\beta}, \sigma|y) d\sigma, \tag{2.8}$$

In a similar manner, an estimate of σ, denoted as $\hat{\sigma}$, can be obtained using the marginal posterior distribution of σ,

$$\pi(\sigma|y) = \int \pi(\boldsymbol{\beta}, \sigma|y) d\boldsymbol{\beta}, \tag{2.9}$$

Although a standard conjugate prior distribution is not available for the mode regression formulation, Markov Chain Monte Carlo (MCMC) methods may be used for extracting the posterior distributions of both $\boldsymbol{\beta}$ and σ.

2.3. Estimation of Covariance Matrix of Classical Estimates

Under the classical approaches of Lee (1989, 1993) and Kemp and Santos Silva (2012), the covariance matrix, $cov\{\hat{\boldsymbol{\beta}}\}$ of the classical estimator $\hat{\boldsymbol{\beta}}$ and its inverse are often required but difficult to estimate or compute numerically, especially under small or moderate samples. A by-product of the proposed Bayesian approach is that using the MCMC posterior sample leads to a natural and efficient estimation of $cov\{\hat{\boldsymbol{\beta}}\}$ and other asymptotic quantities of $\hat{\boldsymbol{\beta}}$.

In fact, a MCMC scheme constructs a Markov chain whose equilibrium distribution is the joint posterior, $p(\boldsymbol{\beta}|data)$. After running the Markov chain for a burn-in period, one obtains samples from the limiting distribution, provided that the Markov chain has converged. Given that the chain has converged, the frequency of appearance of the parameters in the Markov chain represents their posterior distribution. An informative full density distribution of the model parameters is readily obtained rather than a single point estimate as in the classical approach.

When a Markov chain, S, is drawn from the posterior distribution, $p(\boldsymbol{\beta}|data)$: $S = (\boldsymbol{\beta}^{(1)}, \boldsymbol{\beta}^{(2)}, ..., \boldsymbol{\beta}^{(N)})$, where N is the number of draws after burn-in, a consistent estimate of the inverse of the covariance matrix $cov\{\hat{\boldsymbol{\beta}}\}$ can be obtained by multiplying by N the variance-covariance matrix of this MCMC sequence (Chernozhukov and Hong 2003).

2.4. Prior Section and Proper posteriors

In this section first we demonstrate that almost all priors for $(\boldsymbol{\beta}, \sigma)$ could be used and yield a proper joint posterior. In fact we have the following theorem.

Thm 2.1. *Given the mode regression (2.4) and the 'working' likelihood (2.6), if the joint prior distribution $\pi(\boldsymbol{\beta}, \sigma)$ follows one of the following three choices:*

(1) $\pi(\boldsymbol{\beta}, \sigma) \propto 1$ (totally non-informative prior)

(2) $\pi(\boldsymbol{\beta}, \sigma) = \pi(\boldsymbol{\beta})\,\pi(\sigma|\boldsymbol{\beta})$ and one of $\pi(\boldsymbol{\beta})$ and $\pi(\sigma|\boldsymbol{\beta}) \propto 1$ and the other is a proper prior,

(3) $\pi(\boldsymbol{\beta}, \sigma) = \pi(\boldsymbol{\beta})\,\pi(\sigma|\boldsymbol{\beta})$ and both $\pi(\boldsymbol{\beta})$ and $\pi(\sigma|\boldsymbol{\beta})$ are proper priors,

then the posterior distribution of $\boldsymbol{\beta}$ and σ, $\pi(\boldsymbol{\beta}, \sigma|\boldsymbol{y})$, will be a proper distribution. In other words

$$0 < \int \pi(\boldsymbol{\beta}, \sigma|\boldsymbol{y})\, d\boldsymbol{\beta}\, d\sigma < \infty,$$

or, equivalently,

$$0 < \int L(\boldsymbol{y}|\boldsymbol{\beta}, \sigma)\, \pi(\boldsymbol{\beta}, \sigma)\, d\boldsymbol{\beta}\, d\sigma < \infty.$$

The proof can be found in the Appendix.

In practice one usually assumes that the components of $\boldsymbol{\beta}$ have independent prior distributions which is a special case of the above theorem.

2.5. One practical selection of prior on σ

If the conditional distribution is strictly unimodal and symmetric or if the regressors affect only the location of the distribution, then a consistent estimate of the mode can be obtained with a fixed σ (Lee (1989)). In practice, however, data with such characteristics is relatively rare. In addition, in such cases the added value of mode regression is rather limited as the mode coincides with the mean and the median. To extend mode regression to more interesting applications σ must be allowed to approach zero as the sample size goes to infinity.

A suitable prior distribution for σ would be one with a positive support. To this end it is proposed to use either a Uniform(w_1, w_2) or a Gamma distribution with mean w_i, where, in both cases w_i can be determined using one of the following options, commonly used in bandwidth selection methods for kernel density estimation:

- The empirical rule, which states that, given a symmetric distribution, approximately 99.7% of the data values fall within three standard deviations (sd) of the mean, therefore, $w_i = 3sd$;
- Variations of Silverman's plug-in estimate for the bandwidth (Silverman (1986)), in which $w_i = 1.3643\delta n^{-0.2}[min(\widehat{sd}, IQR/1.349)]$, where, IQR is the sample inter quantile range and $\delta = 1.3510$ for a uniform kernel. To cover data with large number of outliers $IQR/1.349$ can be replaced by $1.4826MAD$, where MAD is the median absolute deviation.

Alternatively, as the next section demonstrates, a more flexible model can be developed by relaxing the distributional assumption on the prior for σ using a Dirichlet process prior. This leads to a flexible nonparametric mixture model. The method is nonparametric in the sense that it is not assumed that the prior belongs to any fixed class of distributions.

2.6. Nonparametric Bayesian method

In this section, we formulate a nonparametric Bayesian mode regression model to avoid critical dependence on the mode uniform distribution assumption thus to address the issue of misspecification that may arise under the parametric Bayesian method.

A density $f(\cdot)$ on $\mathbb{R}^+$ is non-increasing if and only if there exists a distribution function G such that $f(x|G) = \int \sigma^{-1} I_{[0<x<\sigma]} dG(\sigma)$ (Feller 1971). Therefore, any unknown density $f(\cdot)$ (with mode θ), symmetric or not, can be represented as a scale mixture of symmetric uniform distributions, that is

$$f(x|\theta, G) = \int \frac{1}{2\sigma} I_{[-\sigma<x-\theta<\sigma]} dG(\sigma), \tag{2.10}$$

where G is the mixing distribution supported on $\mathbb{R}^+$.

Then, a nonparametric Bayesian mode regression model can be expressed in the hierarchical form

$$\begin{aligned} &y_i|\boldsymbol{\beta}, \sigma_i \overset{ind}{\sim} f(y_i - x_i'\boldsymbol{\beta}; \sigma_i), i = 1 \cdots n \\ &\sigma_i|G \overset{iid}{\sim} G, i = 1 \cdots n \\ &G|M, d \sim DP(M, G_0(\cdot, d)) \\ &\boldsymbol{\beta}, M, d \sim p(\boldsymbol{\beta}), p(M), p(d), \end{aligned} \tag{2.11}$$

where, G is the mixing distribution, with base distribution G_0 and concentration parameter M and $f(y_i - x_i'\boldsymbol{\beta}; \sigma_i) = \frac{1}{2\sigma} I_{[-\sigma<y_i - x_i'\boldsymbol{\beta}<\sigma]}$ is the density of a uniform distribution on $(-\sigma, \sigma)$.

We take a uniform distribution as the base distribution, G_0, uniform prior for M and we choose non-informative Normal priors for all the components of $\boldsymbol{\beta}$.

2.7. Empirical Likelihood based Bayesian Method

In addition to parametric and nonparametric likelihood, an empirical likelihood based method could be an alternative for Bayesian mode regression. To derive an empirical likelihood for mode regression we begin with notations and a moment restriction. Lee (1993) generalized the mode regression estimator of Lee (1989), $\hat{\boldsymbol{\beta}} = argmin_{\boldsymbol{\beta}} E\{L(Y - \boldsymbol{x}'\boldsymbol{\beta})\}$, by using the triangular kernel $L(Y; \mu) = \{(\sigma^2 - (Y - \mu)^2) I[|Y - \mu| < \sigma]\}$.

Therefore, the moment restriction for the empirical likelihood can be obtained by the derivative

$\frac{\partial}{\partial\mu}L(Y;\mu) = 2(Y-\mu)I[|Y-\mu| < \sigma]$. Let $l(Y;\mu)$ be the 'derivative' of $L(.;\mu)$ with respect to μ, then the mode, μ, of Y satisfies the moment restriction $E(l(Y;\mu)) = 0$, where $l(Y;\mu) = (Y-\mu)\,I(|Y-\mu| < \sigma)$.

Under an empirical likelihood for mode regression $\mu = \boldsymbol{x}'\boldsymbol{\beta}$, thus for any proposed $\boldsymbol{\beta}$ to estimate the true p dimensional $\boldsymbol{\beta}_0$ via empirical likelihood, we use the vector estimating functions $g(X,Y,\boldsymbol{\beta})$ with component $g_j(X,Y,\boldsymbol{\beta}) = l(Y;\boldsymbol{\beta}'X)\,X_j$ for $j = 1,..,p$. Then, the profile empirical likelihood ratio is given by

$$\Re(\boldsymbol{\beta}) = \max\{\prod_{i=1}^{n}(n\,p_i)|\sum_{i=1}^{n}p_i\,g(X_i,Y_i,\boldsymbol{\beta}) = 0,\ p_i \geq 0,\ \sum_{i=1}^{n}p_i = 1\}.$$

By a standard Lagrange multiplier argument we have

$$\Re(\boldsymbol{\beta}) = \prod_{i=1}^{n}\{n\,p_i(\boldsymbol{\beta})\}, \tag{2.12}$$

with the weights $p_i(\boldsymbol{\beta}) = \frac{1}{n(1+\hat{\lambda}(\boldsymbol{\beta})'g(X_i,Y_i,\boldsymbol{\beta}))}$, where the Lagrange multiplier $\hat{\lambda}(\boldsymbol{\beta})$ is the solution of λ to the following equation

$$\sum_{i=1}^{n}\frac{g(X_i,Y_i,\boldsymbol{\beta})}{1+\lambda^T\,g(X_i,Y_i,\boldsymbol{\beta})} = 0. \tag{2.13}$$

According to Qin and Lawless (1994), among others, the existence and uniqueness of $\hat{\lambda}(\boldsymbol{\beta})$ are guaranteed when the following two conditions are satisfied: (1) zero belongs the convex hull of $\{g(X_i,Y_i,\boldsymbol{\beta}),\ i = 1,...,n\}$ and (2) the matrix $\sum_{i=1}^{n}\{g(X_i,Y_i,\boldsymbol{\beta})g(X_i,Y_i,\boldsymbol{\beta})'\}$ is positive definite.

Under Bayesian inference we consider the empirical likelihood function $\Re(\boldsymbol{\beta})/n^n = \prod_{i=1}^{n}\{p_i(\boldsymbol{\beta})\}$, which can be combined with a prior specification $\pi(\boldsymbol{\beta})$ on the parameter $\boldsymbol{\beta}$ to obtain the posterior distribution

$$\pi(\boldsymbol{\beta}|data) \propto \pi(\boldsymbol{\beta})\,\Re(\boldsymbol{\beta}).$$

2.8. Asymptotic Properties of Bayesian Empirical Likelihood

Before establishing the asymptotic normality of the empirical likelihood-based Bayesian mode regression parameter estimates, the consistency of the empirical likelihood estimator must be established, which is a necessary condition for the asymptotic normality of the posterior. Since the criterion function $g(X,Y,\boldsymbol{\beta})$ results in a non-smooth estimating equations, a similar method to

the one used by Molanes Lopez et al. (2009), among others, is employed to derive the asymptotic results.

Let $\hat{\boldsymbol{\beta}} = argmax_{\boldsymbol{\beta}} \Re(\boldsymbol{\beta})$ be the maximum empirical likelihood estimator (MELE) in a compact set of parameter space which contains the true parameter $\boldsymbol{\beta}_0$. Then note that the criterion function $g(X, Y, \boldsymbol{\beta})$ can be regarded as a special case of M-estimators as discussed in Chapter 5 of Van der Vaart (1998) and satisfies the conditions of theorem 5.7 in the book. Under some regular conditions imposed on the marginal distribution of X and on the conditional distribution of Y given X, such as uniformly continuous and bounded, and since both $E\{g(X, Y, \boldsymbol{\beta})\}$ and $E\{g(X, Y, \boldsymbol{\beta})\, g(X, Y, \boldsymbol{\beta})'\} > 0$ are sufficiently smooth in a compact set of parameter space, which contains $\boldsymbol{\beta}_0$, the consistency condition C_3 of Molanes Lopez et al. (2009) holds. Then the consistency of empirical likelihood estimates is established. Specifically, a rigorous statement of the conditions and theorem is as follows:

Assumption 1. There exists a neighborhood $\mathcal{N}$ of $\boldsymbol{\beta}_0$ such that $P(\Re(\boldsymbol{\beta}) > 0) \to 1$ for any $\boldsymbol{\beta} \in \mathcal{N}$, as $n \to \infty$.

Assumption 2. The distribution function G_X of X has bounded support $\mathcal{X}$.

Assumption 3. The conditional distribution $F_X(t)$ of Y given X is twice continuously differentiable in t for all $X \in \mathcal{X}$.

Assumption 4. At any $X \in \mathcal{X}$, the conditional density function $F'_X(t) = f_X(t) > 0$ for t in a neighborhood of $\boldsymbol{\beta}'_0 X$.

Assumption 5. $E\{g(X, Y, \boldsymbol{\beta}_0)\, g(X, Y, \boldsymbol{\beta}_0)'\} > 0$ is positive definite.

Thm 2.2. *Under Assumptions 1–5, the MELE $\hat{\boldsymbol{\beta}}$ is a consistent estimator of $\boldsymbol{\beta}_0$.*

Assumptions 1-5 are standard conditions in this kind of asymptotic problems. For example, these conditions are basically similar to Assumptions 3.1-3.5 of Yang and He (2012, pp. 1110) for Bayesian empirical likelihood quantile regression. Assumption 1 is to guarantee that the interior of the convex hull of $\{g(X_i, Y_i, \boldsymbol{\beta}) : i = 1, \cdots, n\}$ for $\boldsymbol{\beta} \in \mathcal{N}$ contains the vector of zeros with probability tending to one. Assumption 4 ensures that $\boldsymbol{\beta}_0$ is indeed the unique solution for $Eg(X, Y, \boldsymbol{\beta}) = 0$. The proof of Theorem 2.2 is sketched in the Appendix.

The asymptotic normality of the posterior distribution $\pi(\boldsymbol{\beta}|data)$ could be established using the fact that the empirical log-likelihood ratio for $\boldsymbol{\beta}$ is well approximated by certain quadratics in the

sense of Lemma 6 of Molanes Lopez et al. (2009) so that,

$$\Gamma_n(\boldsymbol{\beta}) \equiv -n^{-1}\sum_{i=1}^{n}\log(1+\hat{\lambda}(\boldsymbol{\beta})' g(X_i,Y_i,\boldsymbol{\beta})) \tag{2.14}$$

$$= -\frac{1}{2}(\boldsymbol{\beta}-\boldsymbol{\beta}_0)' V_{12}' V_{11}^{-1} V_{12}(\boldsymbol{\beta}-\boldsymbol{\beta}_0) + n^{-1/2}(\boldsymbol{\beta}-\boldsymbol{\beta}_0)' V_{12}' V_{11}^{-1} W_n$$
$$- \frac{1}{2} n^{-1} W_n' V_{11}^{-1} W_n + o_P(n^{-1}), \tag{2.15}$$

with matrices $V_{11} = (E\{g_j(X,Y,\boldsymbol{\beta}_0)\, g_k(X,Y,\boldsymbol{\beta}_0)'\})_{j,k=1}^{p}$, $V_{12} = (-\frac{\partial}{\partial\beta_k} E\{g_j(X,Y,\boldsymbol{\beta})\}|_{\boldsymbol{\beta}=\boldsymbol{\beta}_0})_{j,k=1}^{p}$, and vector $W_n = n^{-1/2}\sum_{i=1}^{n} g(X_i,Y_i,\boldsymbol{\beta}_0))$.

Specifically, we make one more assumption on the prior specification $\pi(\boldsymbol{\beta})$.

Assumption 6. $\log\{\pi(\boldsymbol{\beta})\}$ has bounded first derivative in a neighborhood of $\boldsymbol{\beta}_0$.

Then from $\log \mathfrak{R}(\boldsymbol{\beta}) = n\Gamma_n(\boldsymbol{\beta})$ we have

Thm 2.3. *Under Assumptions 1-6, the posterior density of $\boldsymbol{\beta}$ has the following expansion on any sequence of sets $\{\boldsymbol{\beta} : \boldsymbol{\beta}-\boldsymbol{\beta}_0 = O(n^{-1/2})\}$,*

$$\pi(\boldsymbol{\beta}|data) = \pi(\boldsymbol{\beta})\,\mathfrak{R}(\boldsymbol{\beta}) \propto \exp\{-\frac{1}{2}(\boldsymbol{\beta}-\hat{\boldsymbol{\beta}})' I_n (\boldsymbol{\beta}-\hat{\boldsymbol{\beta}}) + Q_n\} \tag{2.16}$$

with $I_n = nV_{12}' V_{11}^{-1} V_{12}$ and empirical likelihood estimate $\hat{\boldsymbol{\beta}}$ and $Q_n = o_p(1)$. When I_n is positive definite, we have $I_n^{1/2}(\boldsymbol{\beta}-\hat{\boldsymbol{\beta}})$ converging in distribution to $N(0,I)$.

The proof of Theorem 2.3 is sketched in the Appendix.

We finally remark that, as similarly remarked for quantile regression by Remark 3.2 of Yang and He (2012, pp. 1110), the posterior will be improper for flat priors on $\boldsymbol{\beta}$ in the Bayesian empirical likelihood approach for our mode regression, and therefore we should avoid using flat priors on $\boldsymbol{\beta}$.

In the case of the prior distribution shrinking with n, we may use $\pi_n(\boldsymbol{\beta})$ satisfying condition similar to Assumption 3.7 of Yang and He (2012) as priors for our mode regression; see Theorem 3.3 of Yang and He (2012) for details.

3. NUMERICAL EXPERIMENTS

In this section we demonstrate our approach to Bayesian mode regression through two simulated and one real examples. For the real example we consider a dataset which investigates how factors

such as gender, age, consumption of alcohol, consumption of fruit and vegetables and smoking can affect the body mass index (BMI).

3.1. Simulation Example 1

We consider a simulated data from the model

$$y_i = \beta_0 + \beta_1 x_i + \epsilon_i, \tag{3.1}$$

where $x_i \sim N(0,1)$, $i = 1, ..., n$ for $n = 50, 100, 200$ and $\boldsymbol{\beta} = (1, 2)$. The following three specifications were considered for the model error ϵ:

- Case 1: the standard Normal distribution, $\epsilon_i \sim N(0,1)$ - a symmetric error distribution.
- Case 2: a Fisher's Z distribution, $\epsilon_i \sim 1/2logZ$ with $Z \sim F_{2,2}$ - a skewed error distribution.
- Case 3: a Normal distribution with normally distributed outliers (contaminants) centred at twice the distance between the true mode and the 99^{th} percentile of the original Normal distribution and accounting for 20% of the total data points, $\epsilon_i \sim 0.80N(0, \frac{1}{4}) + 0.20N(2.5, \frac{1}{4})$ (Hedges and Shah (2003)) - an asymmetric error distribution.

We fit parametric Bayesian mode regression (labeled PBMR) for all the cases above. Then for demonstration and comparison purposes we fit empirical likelihood based Bayesian mode regression (labeled ELBMR) for case 2 and nonparametric Bayesian mode regression (labeled NBMR) for case 3.

For the PBMR and ELBMR models, independent Normal distributions were used as priors of each component of $\boldsymbol{\beta}$, where the mean and standard derivation of the Normal prior are given by the classical estimator of Lee (1989, 1993) and its estimated standard error respectively. Realisations were simulated from the posterior distributions by means of a single-component Metropolis-Hastings algorithm. Each of the parameters was updated using a random-walk Metropolis algorithm with a Gaussian proposal density centred at the current state of the chain. The variance of the proposal density was chosen to provide an acceptance rate close to the optimal acceptance rate as defined in Roberts and Rosenthal (2001). Convergence was assessed using time series plots and the R package

Table 1: Simulation Example 1: True parameter values (T.V.) and their posterior means, standard deviations (S.D.) and 95% credible intervals (C.I)

		PBMR						ELBMR		NBMR	
		Normal		Skewed		Asymmetric		Skewed		Asymmetric	
n		β_0	β_1	β_0	β_1	β_0	β_1	β_0	β_1	β_0	β_1
50	T.V	1	2	1	2	1	2	1	2	1	2
	Mean	0.92	2.00	1.07	2.01	0.96	2.02	1.01	2.00	1.09	1.94
	S.D.	0.78	0.77	0.78	0.49	0.34	0.24	0.01	0.01	0.24	0.19
	95%CI	(-0.6,2.1)	(0.5,3.3)	(-0.3,2.6)	(1.2,3.1)	(0.4,1.7)	(1.6,2.5)	(0.99,1.02)	(1.99,2.01)	(0.7,1.5)	(1.5,2.3)
100	T.V	1	2	1	2	1	2	1	2	1	2
	Mean	1.01	2.10	0.95	1.89	1.06	1.94	1.01	2.00	1.06	2.00
	S.D.	0.18	0.25	0.52	0.37	0.98	0.76	0.01	2	0.14	0.12
	95%CI	(0.6,1.3)	(1.6,2.6)	(0.0,1.9)	(1.2,2.6)	(-0.7,2.9)	(0.5,3.3)	(0.99,1.02)	(1.99,2.01)	(0.8,1.3)	(1.8,2.2)
200	T.V	1	2	1	2	1	2	1	2	1	2
	Mean	1.26	1.99	1.00	1.99	1.06	1.96	1.01	2.00	1.04	1.91
	S.D.	0.86	0.52	1.29	0.75	0.82	0.42	0.01	0.01	0.07	0.06
	95%CI	(-0.5,2.8)	(0.9,3.0)	(-1.3,3.5)	(0.6,3.3)	(-0.4,2.6)	(1.2,2.7)	(0.99,1.02)	(1.99,2.01)	(0.92,1.19)	(1.78,2.03)

boa (Smith (2007)). The estimates are posterior means using 10,000 iterations of the MCMC sampler (after 10,000 burn-in iterations).

The estimates for the NBMR model were obtained by fitting a truncated Dirichlet Process (DP) mixture model, which leads to a computationally straightforward approximation and can be easily implemented in the freely available WinBUGS software. Two parallel chains of equal length with different initial values were run for the model. The results were based on 10,000 iterations which followed a burn-in period of 40,000 for each chain.

Table 1 compares the posterior means with the true values of β_0 and β_1 and also gives standard deviations and 95% credible intervals for each of the models considered in this example.

The results of the analysis indicate that the PBRM works well, as all the absolute biases for the estimated parameters turn out to be in the range [0.01, 0.26]. Furthermore, under both ELBMR and NBRM, the true values for both β_0 and β_1 are recovered successfully indicating that the methods also work well. However, it should be noted that the standard deviations for both parameters are smaller than in the PBMR, giving shorter confidence intervals.

The MCMC sampler for the regression parameters $\widehat{\beta_0}$ and $\widehat{\beta_1}$ can be used to obtain the empirical samples from the joint posterior distributions of the PBMR parameters. These samples can be used to obtain a consistent estimator of the covariance or correlation structure of the parameter estimators, which is difficult to estimate under the classical approach. For example in case (a),

with sample size n=100, we have

$$\widehat{Cov}\begin{pmatrix}\hat{\beta}_0\\ \hat{\beta}_1\end{pmatrix} = \begin{pmatrix}3 & -1\\ -1 & 6\end{pmatrix}.$$

3.2. Simulation Example 2

In this section we present the results of a second simulation example with the aim of comparing the performance of our approach with the classical mode regression approach. Specifically, we replicate the simulation study in Kemp and Santos Silva (2012), but only for a sample of size 250, and compare their results with the results obtained under our Bayesian mode regression approach.

Simulation data are generated by the simple linear model

$$y_i = \beta_0 + \beta_1 x_i + (1 + v x_i)\epsilon_i, \tag{3.2}$$

where x_i are generated from a $\chi^2_{(3)}$ distribution, scaled to have variance 1, and ϵ_i are generated as independent draws from a re-scaled log-gamma random variable,

$$\epsilon_i = -\lambda\, ln(Z_i), \tag{3.3}$$

where Z follows a gamma distribution with mean 1 and scale parameter $\frac{1}{\alpha}$, to ensure that ϵ_i has zero mode. Furthermore, we set $\lambda = [(1 + 2E(x_i)v + E(x_i^2)v^2)\psi(\alpha)]$ [1] to ensure that the unconditional variance of the error $(1 + vx_i)$ is equal to one.

The study was performed for $\alpha \in \{0.05, 5\}$ and for $v \in \{0, 2\}$. Table 1 compares the 95% Bayesian credible intervals (BCI) for the estimates obtained under PBMR and NBMR with the 95% classical confidence intervals (CI) for the estimates under the two classical mode regression models: Mode 1.6 and Mode 0.8. Mode 1.6 and Mode 0.8 correspond to $k = 1.6$ and $k = 0.8$ respectively in the bandwidth selection rule, bandwidth=$k\, mad\, n^{-0.143}$, where mad is the median of the absolute deviation from the median of ordinary least squares regression residuals.

The results of the analysis suggest that the Bayesian mode regression estimates are strong competitors of the classical mode regression estimates since in almost all the examples both PBMR

[1] $\psi(\cdot)$ is the trigamma function

Table 2: Simulation Example 2: Comparison between Classical and Bayesian approach for mode regression

			PBMR	NBMR	Mode 1.6	Mode 0.8
α	n		95% BCI	95% BCI	95% CI	95% CI
5.00	0	β_0	(-0.37,0.29)	(-0.21,0.36)	(-0.31, 0.41)	(-0.69, 0.75)
		β_1	(0.82,1.28)	(0.89,1.32)	(0.77, 1.24)	(0.56,1.45)
	2	β_0	(-0.06,0.07)	(-0.03,0.21)	(-0.15,0.23)	(-0.25,0.29)
		β_1	(0.99,1.14)	(0.80,1.22)	(0.63,1.37)	(0.48,1.53)
0.05	0	β_0	(0.00, 0.14)	(-0.03,0.07)	(0.12,0.42)	(-0.09,0.35)
		β_1	(0.95,1.13)	(0.95,1.06)	(0.90,1.11)	(0.87,1.17)
	2	β_0	(0.02,0.08)	(0.04,0.09)	(0.09,0.29)	(0.01,0.21)
		β_1	(0.99,1.08)	(0.97,1.04)	(0.91,1.19)	(0.85,1.19)

and NBMR estimators outperform the two classical estimators.

Finally, as also evident from Kemp and Santos Silva (2012), the selection of the value/prior for σ plays an important role on the precision of the parameters.

3.3. Factors Affecting the Body Mass Index (BMI)

Following the introduction of the BMI example in Section 1, the proposed methodology was applied to investigate the research question: "What is the effect of factors such as gender, age, consumption of alcohol, consumption of fruit and vegetables and smoking on the typical body mass index (BMI)?"

A person's typical BMI was modelled as a function of the person's age, age_i, the total units of alcohol consumed per week, $alcohol_i$, the portion of fruit and vegetables consumed the previous day, $fruit\&veg_i$ the person's cigarette smoking status, $smoking_i$ (1= Non-smoker, 2= Light smokers, under 10 a day, 3= Moderate smokers, 10 to under 20 a day, 4=Heavy smokers, 20 or more a day), and of a gender indicator, $male_i$ (1=male, 0=female):

$$bmi_i = \beta_0 + \beta_1 age_i + \beta_2 alcohol_i + \beta_3 fruit\&veg_i + \beta_4 smoking_i + \beta_5 gender_i + \epsilon_i \quad (3.4)$$

The BMI range is from 15.9 to 56.0 (range =40.1) indicating a significant disparity between high and low BMI scores. The average BMI is 27.75 with standard deviation of 5.13. The high

levels for range and standard deviation suggest the presence of outliers which cause the mean to be pulled in the direction of the tail. As a consequence, the mean, median, and mode do not coincide and it can be easily concluded that the distribution of the data is positively skewed. Figure **??** in section 1 demonstrates the density of BMI for the total, males and females, verifying that all three distributions are positively skewed. The mode represents the most typical value and is the value at the peak of the distribution. Even though, mean regression and quantile regression could have been applied to model BMI these methods cannot reveal any information about the mode, or about the effect of the covariates on the most typical case.

Table 3 presents the estimation results obtained with the traditional mean, quantile regressions and with the proposed mode regression. The analysis was performed for the total of responders but also for males and females separately. For the mode regression, an independent improper uniform prior was chosen for all the components of $\boldsymbol{\beta}$ and a gamma prior with mean $3sd(\mathbf{bmi})$ for σ. Realisations were simulated from the posterior distributions by means of a single-component Metropolis-Hastings algorithm. Each of the parameters was updated using a random-walk Metropolis algorithm with a Gaussian proposal density centred at the current state of the chain. The estimates are posterior means using 10,000 iterations of the MCMC sampler (after 10,000 burn-in iterations).

As expected, mean regression indicates that on average the BMI is lower for women than for men but, as indicated by quantile regression, the effect of gender differs significantly at different quantile levels. More specifically, at the 25% level, the BMI of women is around 1.36 units lower than the corresponding BMI for men but this gap is smaller for the median case (0.75) and decreases further at the 75% quantile level (0.29). Mode regression reveals that the gender differential in the most typical BMI is lower than both the mean and the median, since as opposed to the other statistics, mode is not influenced by the extreme observations. According to the results, the typical BMI for women is 0.27 units lower than the corresponding BMI for men.

The effect of fruit and vegetables is not significantly different from zero for mean and quantile regression at the 25% level and at the median. However, at the 75% level the consumption of additional fruit and vegetables has a negative effect on the typical BMI (-0.11). Similar results are obtained for both males and females, however, under quantile regression for males the effect is not significant at any quantile levels. However, under mode regression it seems that the consumption of

Table 3: BMI dataset: Estimation results for mean, quantile and mode regression

	Mean Regression		Quantile Regression 0.25		Quantile Regression 0.50		Quantile Regression 0.75		Parametric Bayesian Mode Regression	
Total (n=4,138)										
Variable	Estimate	S.E.	Estimate	S.E.	Estimate	S.E.	Estimate	S.E.	Estimate	S.D.
const	25.74	0.33	21.32	0.29	24.04	0.37	28.42	0.53	34.44	10.6
age	0.05	0.004	0.05	0.004	0.06	0.005	0.05	0.01	-0.01	0.16
alcohol	-0.01	0.004	0.004	0.001	0.001	0.004	-0.004	0.001	-0.01	0.05
fruit&veg	-0.07	0.03	-0.002	0.02	-0.06	0.02	-0.11	0.04	-0.24	0.54
smoking	-0.37	0.13	-0.40	0.11	-0.34	0.17	-0.36	0.20	-1.16	3.34
male	0.42	0.16	1.36	0.15	0.75	0.17	0.29	0.24	0.27	6.15
Males (n=1,814)										
const	25.78	0.45	22.81	0.44	24.85	0.47	27.99	0.70	34.7	9.70
age	0.05	0.006	0.05	0.006	0.06	0.01	0.05	0.01	-0.003	0.15
alcohol	0.01	0.004	0.01	0.005	0.01	0.01	0.01	0.01	-0.01	0.06
fruit&veg	-0.03	0.04	-0.02	0.04	-0.03	0.04	-0.02	0.06	-0.31	0.68
smoking	-0.43	0.19	-0.57	0.20	-0.48	0.24	-0.51	0.33	-1.47	3.53
Females (n=2,324)										
const	26.05	0.47	21.36	0.40	23.76	0.47	28.51	0.83	33.8	10.25
age	0.04	0.006	0.05	0.006	0.06	0.01	0.05	0.01	0.01	0.16
alcohol	-0.03	0.01	-0.01	0.01	-0.01	0.01	-0.02	0.01	-0.04	0.01
fruit&veg	-0.1	0.04	0.01	0.03	-0.07	0.04	-0.15	0.06	-0.17	0.53
smoking	-0.27	0.20	-0.29	0.18	-0.19	0.27	-0.29	0.37	-1.32	3.73

an additional unit of fruit and vegetables is negatively correlated with the typical BMI level. The negative effect for males is almost twice as high as for females. This results imply that, typically, eating more fruit and vegetables can contribute to lowering BMI levels; thus losing weight.

Furthermore, the results of the analysis suggest that heavier smoking is also negatively correlated with the BMI under all three methods. However, under mode regression the negative effect of heavier smoking on the typical BMI is 3 times higher as compared to the effect on the mean BMI and at different quantile levels.

Finally, under all three methods, the effect of age and alcohol cannot be considered as significantly different from zero for the total, but also for males and females separately.

In conclusion, the results indicate that mode regression is a useful statistical technique, especially when analysing data with outliers. In this example, even though the overall effect of covariates on the response variable was similar under the three regression methods, the marginal effects of the covariates were often different, justifying the usefulness of mode regression as an alternative analysis tool.

4. CONCLUSIONS

Identifying the typical value or pattern could be one of the most efficient statistical methods of data analysis, in particular, for big data analysis. In this paper a novel Bayesian mode regression framework has been presented which includes three approaches: a parametric method, a nonparametric method and an empirical likelihood-based method. It should be noted that, in the area of mode regression, there is no literature from a Bayesian perspective. The paper demonstrates that the estimates are consistent and asymptotically Normal under fairly standard conditions and even under misspecification of the likelihood function. The numerical studies suggest that the proposed Bayesian mode regression estimates are strong competitors to the classical mode regression estimates.

APPENDIX: PROOFS OF THEOREMS

Proof of theorem 2.1

The γth moments of marginal posterior distribution of $\boldsymbol{\beta}$ is given by

$$E[|\boldsymbol{\beta}|^{\gamma}|\sigma,\boldsymbol{y}] = \int \frac{1}{(2\sigma)^n} \prod_{i=1}^{n} I[|y_i - x_i'\boldsymbol{\beta}| < \sigma]\, \pi(\boldsymbol{\beta},\sigma) d\boldsymbol{\beta}\, d\sigma.$$

Noting that $\prod_{i=1}^{n} I[|y_i - x_i'\boldsymbol{\beta}| < \sigma]$ provides joint bands for all components β_j $(j = 0, 1, ..., p)$ of $\boldsymbol{\beta}$. Let us say $0 < |\beta_j| < B_j < \infty$ $(j = 0, 1, ..., p)$, even if some of $|y_i - x_i'\boldsymbol{\beta}| < \sigma$ are true and some are not. Therefore,

$$E[|\boldsymbol{\beta}|^{\gamma}|\sigma,\boldsymbol{y}] = \int \frac{1}{(2\sigma)^n} d\sigma \int_{-B_0}^{B_0} \int_{-B_1}^{B_1} ... \int_{-B_p}^{B_p} \prod_{j=0}^{p} |\beta_j|^{r_j}\, \pi(\boldsymbol{\beta},\sigma)\, d\boldsymbol{\beta},$$

which is clearly finite. Similarly, for the γth moment of marginal posterior of σ with $\gamma < n$ is defined as $E[|\sigma|^{\gamma}|\boldsymbol{\beta},\boldsymbol{y}]$, and can be provided finite in the same way.

Proof of theorem 2.2

We will show Theorem 2.2 by applying a generic consistency lemma, Lemma 4.1, of Lu et al. (2007). For convenience of statement, we define $R_n(\lambda,\boldsymbol{\beta}) \equiv n^{-1}\sum_{i=1}^{n} \log(1 + \lambda' g(X_i, Y_i, \boldsymbol{\beta}))$ and $R(\lambda,\boldsymbol{\beta}) \equiv E\{\log(1 + \lambda' g(X_i, Y_i, \boldsymbol{\beta}))\}$. Then note that $R_n(\hat{\lambda}(\boldsymbol{\beta}),\boldsymbol{\beta}) = -\Gamma_n(\boldsymbol{\beta})$ and $\hat{\boldsymbol{\beta}} = \arg\min_{\boldsymbol{\beta}\in\mathbb{B}} R_n(\hat{\lambda}(\boldsymbol{\beta}),\boldsymbol{\beta})$, where $\hat{\lambda}(\boldsymbol{\beta})$ and $\Gamma_n(\boldsymbol{\beta})$ are defined in (2.13) and (2.14), respectively, and $\mathbb{B}$ is a compact subset of $\mathbb{R}^p$ containing the true parameter vector $\boldsymbol{\beta}_0$ as an interior point. Further, we denote by $H_n(\lambda,\boldsymbol{\beta})$ for the left-hand side of (2.13) divided by n, that is $H_n(\lambda,\boldsymbol{\beta}) \equiv n^{-1}\sum_{i=1}^{n}\{g(X_i, Y_i, \boldsymbol{\beta})/[1 + \lambda' g(X_i, Y_i, \boldsymbol{\beta})]\}$, and hence for any $\boldsymbol{\beta} \in \mathbb{B}$, $\hat{\lambda}(\boldsymbol{\beta})$

is the solution of λ to the equation $H_n(\lambda, \boldsymbol{\beta}) = 0$.

We will need the following lemma on the continuity for the quantities related.

Lemma .1. *Under Assumptions 2 and 3, we have the following results:*

(L1) $E\{g(X, Y, \boldsymbol{\beta})\}$ *and* $E\{g(X, Y, \boldsymbol{\beta})g(X, Y, \boldsymbol{\beta})'\}$ *are twice continuously differentiable with respect to* $\boldsymbol{\beta}$.

(L2) There exist p *dimensional compact neighborhoods* C_λ *and* $C_{\boldsymbol{\beta}}$ *around* 0, *in which* $H_0(\lambda, \boldsymbol{\beta}) = E[g(X, Y, \boldsymbol{\beta})/\{1+\lambda' g(X, Y, \boldsymbol{\beta})\}]$ *is twice continuously differentiable in* $\boldsymbol{\beta} \in C_{\boldsymbol{\beta}}$ *and* $\lambda \in C_\lambda$, *and* $E[g(X, Y, \boldsymbol{\beta})g(X, Y, \boldsymbol{\beta})'/\{1+\lambda' g(X, Y, \boldsymbol{\beta})\}]$ *is uniformly continuous with respect to* $\boldsymbol{\beta} \in C_{\boldsymbol{\beta}}$ *and* $\lambda \in C_\lambda$.

The proof of this lemma is similar to that of Lemma A.1 of Yang and He (2012, pp. 1121). We only need to notice $g(X_i, Y_i, \boldsymbol{\beta}) = (Y_i - \boldsymbol{\beta}' X_i) I_{\{|Y_i - \boldsymbol{\beta}' X_i| < \sigma\}} X_i$ and apply Assumptions 2 and 3. As an illustration, we provide the proof for $Eg(X_i, Y_i, \boldsymbol{\beta})$ here. Note that

$$\begin{aligned} Eg(X_i, Y_i, \boldsymbol{\beta}) &= E_X \int (y - \boldsymbol{\beta}' X) I_{\{|y - \boldsymbol{\beta}' X| < \sigma\}} X f_X(y) dy \\ &= E_X \int_{\boldsymbol{\beta}' X - \sigma}^{\boldsymbol{\beta}' X + \sigma} (y - \boldsymbol{\beta}' X) X f_X(y) dy, \end{aligned}$$

where E_X stands for the expectation with respect to the distribution G_X of the random variable X. Then the first order derivative of $Eg(X_i, Y_i, \boldsymbol{\beta})$ with respect to $\boldsymbol{\beta}$, through simple algebraic calculations, is

$$\frac{\partial Eg(X_i, Y_i, \boldsymbol{\beta})}{\partial \boldsymbol{\beta}} = E_X\{\sigma X(f_X(\boldsymbol{\beta}' X + \sigma) - f_X(\boldsymbol{\beta}' X - \sigma)) - X X'(F_X(\boldsymbol{\beta}' X + \sigma) - F_X(\boldsymbol{\beta}' X - \sigma))\}.$$

Now by Assumptions 2 and 3, clearly $\frac{\partial Eg(X_i, Y_i, \boldsymbol{\beta})}{\partial \boldsymbol{\beta}}$ is further differentiable with respect to $\boldsymbol{\beta}$. The remaining parts of this lemma can be proved similarly with details omitted. ‡

We further define $\lambda_0(\boldsymbol{\beta})$ to be the solution of λ to the equation $H(\lambda, \boldsymbol{\beta}) \equiv E\{g(X_i, Y_i, \boldsymbol{\beta})/[1+\lambda' g(X_i, Y_i, \boldsymbol{\beta})]\} = 0$. By Lemma .1, Assumption 5 and the implicit function theorem, $\lambda_0(\boldsymbol{\beta})$ uniquely exists in the neighbourhood C_λ of $\mathbf{0} \in \mathbb{R}^p$. By this uniqueness, as $Eg(X, Y, \boldsymbol{\beta}_0) = 0$, we have $\lambda_0(\boldsymbol{\beta}_0) = 0$. Therefore it follows that $R(\lambda_0(\boldsymbol{\beta}_0), \boldsymbol{\beta}_0) = E\{\log(1 + (\lambda_0(\boldsymbol{\beta}_0))' g(X_i, Y_i, \boldsymbol{\beta}_0))\} = 0$. Note that under Assumptions 1–5, $\boldsymbol{\beta}_0 = \arg\min_{\boldsymbol{\beta} \in \mathbb{B}} R(\lambda_0(\boldsymbol{\beta}), \boldsymbol{\beta})$.

To show the consistency of $\hat{\boldsymbol{\beta}}$ to $\boldsymbol{\beta}_0$, we will apply a lemma below that is a special case of Lemma 4.1 of Lu et al. (2007). Here we need to define a uniform metric $\|\cdot\|_{\mathbb{B}}$ for the distance of any continuous function $\lambda : \mathbb{B} \mapsto \mathbb{R}^p$ from $\lambda_0(\cdot)$, that is $\|\lambda(\cdot) - \lambda_0(\cdot)\|_{\mathbb{B}} = \sup_{\boldsymbol{\beta} \in \mathbb{B}} \|\lambda(\boldsymbol{\beta}) - \lambda_0(\boldsymbol{\beta})\|$ with $\|\cdot\|$ standing for the Euclidean norm of $\mathbb{R}^p$.

Lemma .2. *Suppose* $\boldsymbol{\beta}_0 \in \mathbb{B}$ *(a compact subset of* $\mathbb{R}^p$*) satisfies* $R(\lambda_0(\boldsymbol{\beta}_0), \boldsymbol{\beta}_0) = \inf_{\boldsymbol{\beta} \in \mathbb{B}} R(\lambda_0(\boldsymbol{\beta}), \boldsymbol{\beta})$, *and that the following hold.*

(i) $R_n\left(\hat{\lambda}(\hat{\boldsymbol{\beta}}), \hat{\boldsymbol{\beta}}\right) \leq \inf_{\boldsymbol{\beta}\in\mathbb{B}} R_n\left(\hat{\lambda}(\boldsymbol{\beta}), \boldsymbol{\beta}\right) + o_P(1)$.

(ii) *For all* $\delta > 0$, *there exists* $\epsilon(\delta) > 0$ *such that*

$$\inf_{\|\boldsymbol{\beta}-\boldsymbol{\beta}_0\|>\delta} R\left(\lambda_0(\boldsymbol{\beta}), \boldsymbol{\beta}\right) \geq R\left(\lambda_0(\boldsymbol{\beta}_0), \boldsymbol{\beta}_0\right) + \epsilon(\delta).$$

(iii) *Uniformly for all* $\boldsymbol{\beta} \in \mathbb{B}$, $R\left(\lambda(\boldsymbol{\beta}), \boldsymbol{\beta}\right)$ *is continuous [with respect to the uniform metric* $\|\cdot\|_{\mathbb{B}}$*] in* $\lambda(\boldsymbol{\beta})$ *at* $\lambda_0(\boldsymbol{\beta})$.

(iv) $\|\hat{\lambda}(\cdot) - \lambda_0(\cdot)\|_{\mathbb{B}} = o_P(1)$.

(v) *For all* $\{\delta_n\}$ *with* $\delta_n = o(1)$,

$$\sup_{\boldsymbol{\beta}\in\mathbb{B}} \sup_{\|\lambda(\boldsymbol{\beta})-\lambda_0(\boldsymbol{\beta})\|_{\mathbb{B}}\leq\delta_n} \left|R_n\left(\lambda(\boldsymbol{\beta}), \boldsymbol{\beta}\right) - R\left(\lambda(\boldsymbol{\beta}), \boldsymbol{\beta}\right)\right| = o_P(1).$$

Then $\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}_0 = o_P(1)$.

The proof of this lemma is omitted; see that of Lemma 4.1 of Lu et al. (2007, pp. 186).

The consistency of $\hat{\boldsymbol{\beta}}$ can be proved by checking the conditions in Lemma .2 step by step: As $\hat{\boldsymbol{\beta}}$ and $\boldsymbol{\beta}_0$ are the minimizers of $R_n(\hat{\lambda}(\boldsymbol{\beta}), \boldsymbol{\beta})$ and $R(\lambda_0(\boldsymbol{\beta}), \boldsymbol{\beta})$, respectively, (i) and (ii) hold obviously. By noting Lemma .1, simple algebraic calculations lead to

$$R(\lambda, \boldsymbol{\beta}) = E_X \int_{\boldsymbol{\beta}'X-\sigma}^{\boldsymbol{\beta}'X+\sigma} \log\{1 + \lambda' X(y - \boldsymbol{\beta}'X)\} f_X(y) dy, \tag{.1}$$

$$H(\lambda_0(\boldsymbol{\beta}), \boldsymbol{\beta}) = E_X \int_{\boldsymbol{\beta}'X-\sigma}^{\boldsymbol{\beta}'X+\sigma} \frac{X(y - \boldsymbol{\beta}'X)}{1 + (\lambda_0(\boldsymbol{\beta}))' X(y - \boldsymbol{\beta}'X)} f_X(y) dy = 0, \tag{.2}$$

and therefore (iii) also holds clearly by the following fact: as $\|\lambda(\cdot) - \lambda_0(\cdot)\|_{\mathbb{B}} \to 0$,

$$
\begin{aligned}
&\sup_{\boldsymbol{\beta}\in\mathbb{B}}|R(\lambda(\boldsymbol{\beta}),\boldsymbol{\beta})-R(\lambda_0(\boldsymbol{\beta}),\boldsymbol{\beta})| \\
&\quad\le \sup_{\boldsymbol{\beta}\in\mathbb{B}}\left|E_X\int_{\boldsymbol{\beta}'X-\sigma}^{\boldsymbol{\beta}'X+\sigma}\left[\log\{1+(\lambda(\boldsymbol{\beta}))'X(y-\boldsymbol{\beta}'X)\}-\log\{1+(\lambda_0(\boldsymbol{\beta}))'X(y-\boldsymbol{\beta}'X)\}\right]f_X(y)dy\right| \\
&\quad\le \sup_{\boldsymbol{\beta}\in\mathbb{B}}\left|E_X\int_{\boldsymbol{\beta}'X-\sigma}^{\boldsymbol{\beta}'X+\sigma}\left[\frac{(\lambda(\boldsymbol{\beta})-\lambda_0(\boldsymbol{\beta}))'X(y-\boldsymbol{\beta}'X)}{1+(\lambda_0(\boldsymbol{\beta}))'X(y-\boldsymbol{\beta}'X)}\right.\right. \\
&\qquad\qquad\left.\left.-\frac{(\lambda(\boldsymbol{\beta})-\lambda_0(\boldsymbol{\beta}))'XX'(y-\boldsymbol{\beta}'X)^2(\lambda(\boldsymbol{\beta})-\lambda_0(\boldsymbol{\beta}))}{[1+(\lambda_0(\boldsymbol{\beta})+\xi(\lambda(\boldsymbol{\beta})-\lambda_0(\boldsymbol{\beta})))'X(y-\boldsymbol{\beta}'X)]^2}\right]f_X(y)dy\right| \\
&\quad\le \|\lambda(\cdot)-\lambda_0(\cdot)\|_{\mathbb{B}}^2\sup_{\boldsymbol{\beta}\in\mathbb{B}}\left|E_X\int_{\boldsymbol{\beta}'X-\sigma}^{\boldsymbol{\beta}'X+\sigma}\left[\frac{\|XX'\|(y-\boldsymbol{\beta}'X)^2}{[1+(\lambda_0(\boldsymbol{\beta}))'X(y-\boldsymbol{\beta}'X)]^2}\right]f_X(y)dy\right|\to 0, \qquad (.3)
\end{aligned}
$$

where $|\xi|<1$, the last inequality follows from equality of (.2), and the last limit is owing to the compactness of $\mathbb{B}$ together with the continuity of the integration part as a function of $\boldsymbol{\beta}$ on the RHS of the last inequality in (.3). (iv) follows from a standard argument of the Z-estimator $\hat{\lambda}(\boldsymbol{\beta})$, which is the solution to $H_n(\lambda,\boldsymbol{\beta})=0$, uniformly converging to $\lambda_0(\boldsymbol{\beta})$, which is the solution to $H(\lambda,\boldsymbol{\beta})=0$, in Chapter 5.1 of Van der Vaart (1998); see also the argument on uniform convergence in the second paragraph on Yang and He (2012, pp. 1124). For (v), letting $\delta_n=o(1)$ and $\|\lambda-\lambda_0\|_{\mathbb{B}}\le\delta_n$, we notice that

$$
\begin{aligned}
&R_n(\lambda(\boldsymbol{\beta}),\boldsymbol{\beta})-R(\lambda(\boldsymbol{\beta}),\boldsymbol{\beta}) \\
&\quad=\{R_n(\lambda(\boldsymbol{\beta}),\boldsymbol{\beta})-R_n(\lambda_0(\boldsymbol{\beta}),\boldsymbol{\beta})\}+\{R_n(\lambda_0(\boldsymbol{\beta}),\boldsymbol{\beta})-R(\lambda_0(\boldsymbol{\beta}),\boldsymbol{\beta})\} \\
&\qquad+\{R(\lambda_0(\boldsymbol{\beta}),\boldsymbol{\beta})-R(\lambda(\boldsymbol{\beta}),\boldsymbol{\beta})\} \\
&\quad=I+II+III,
\end{aligned}
$$

where by (.3) III tends to 0, uniformly for $\boldsymbol{\beta}\in\mathbb{B}$ and with λ satisfying $\|\lambda-\lambda_0\|_{\mathbb{B}}\le\delta_n$. That I tends to 0, uniformly for $\boldsymbol{\beta}\in\mathbb{B}$ and λ with $\|\lambda-\lambda_0\|_{\mathbb{B}}\le\delta_n$, can be proved in the same way as for III, because in fact $E[I]=III$; II can also be proved easily to tend to zero.

Proof of theorem 2.3

Based on the consistency in Theorem 2.2, Theorem 2.3 can be proved similarly to Theorem 3.2 of Yang and He (2012) by noticing the difference of mode regression in this paper from quantile regression in Yang and He (2012). First, under Assumptions 2–4, it is easy to show as done in Lemma A.5 of Yang and He (2012) that

(C1) $\|\sum_{i=1}^n[g(X_i,Y_i,\boldsymbol{\beta})-Eg(X_i,Y_i,\boldsymbol{\beta})]\|=O_p(n^{1/2})$, uniformly in $\boldsymbol{\beta}$ in a $o(1)$-neighborhood of $\boldsymbol{\beta}_0$.

(C2) $\|\sum_{i=1}^n[g(X_i,Y_i,\boldsymbol{\beta})g(X_i,Y_i,\boldsymbol{\beta})'-Eg(X_i,Y_i,\boldsymbol{\beta})g(X_i,Y_i,\boldsymbol{\beta})']\|=o_p(n)$, uniformly in $\boldsymbol{\beta}$ in a $o(1)$-

neighborhood of $\boldsymbol{\beta}_0$.

(C3) $\| \sum_{i=1}^{n} [g(X_i, Y_i, \boldsymbol{\beta}) - Eg(X_i, Y_i, \boldsymbol{\beta}) - g(X_i, Y_i, \boldsymbol{\beta}_0) + Eg(X_i, Y_i, \boldsymbol{\beta}_0)] \| = o_p(n^{-1/2})$, uniformly in $\boldsymbol{\beta}$ for $\boldsymbol{\beta} - \boldsymbol{\beta}_0 = O_p(n^{-1/2})$.

These (C1)-(C3) together with Assumptions 1–5 ensure (2.15) holds true (c.f., Lemma 6 of Molanes Lopez et al. (2009)).

Further, maximizing the main terms on the RHS of (2.15) with respect to $\boldsymbol{\beta}$, we have

$$\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}_0 = n^{-1/2}(V_{12}' V_{11}^{-1} V_{12})^{-1} V_{12}' V_{11}^{-1} W_n + o_P(n^{-1/2}), \tag{.4}$$

where $\hat{\boldsymbol{\beta}}$ is the maximum empirical likelihood estimator of $\boldsymbol{\beta}_0$.

Then it follows from (2.15) and (.4) that

$$\begin{aligned}
\pi(\boldsymbol{\beta}|data) &= \pi(\boldsymbol{\beta})\, \mathfrak{R}(\boldsymbol{\beta}) \\
&= \pi(\boldsymbol{\beta}) \times \exp\Big\{ -\frac{n}{2}(\boldsymbol{\beta} - \boldsymbol{\beta}_0)' V_{12}' V_{11}^{-1} V_{12} (\boldsymbol{\beta} - \boldsymbol{\beta}_0) + n^{1/2} (\boldsymbol{\beta} - \boldsymbol{\beta}_0)' V_{12}' V_{11}^{-1} W_n \\
&\qquad -\frac{1}{2} W_n' V_{11}^{-1} W_n + o_P(1) \Big\} \\
&= \pi(\boldsymbol{\beta}) \times \exp\Big\{ -\frac{n}{2}(\boldsymbol{\beta} - \boldsymbol{\beta}_0)' V_{12}' V_{11}^{-1} V_{12} (\boldsymbol{\beta} - \boldsymbol{\beta}_0) + n (\boldsymbol{\beta} - \boldsymbol{\beta}_0)' V_{12}' V_{11}^{-1} V_{12} (\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}_0) \\
&\qquad -\frac{1}{2} W_n' V_{11}^{-1} W_n + o_P(1) \Big\} \\
&= \pi(\boldsymbol{\beta}) \times \exp\Big\{ -\frac{n}{2}(\boldsymbol{\beta} - \boldsymbol{\beta}_0)' V_{12}' V_{11}^{-1} V_{12} (\boldsymbol{\beta} - 2\hat{\boldsymbol{\beta}} + \boldsymbol{\beta}_0) - \frac{1}{2} W_n' V_{11}^{-1} W_n + o_P(1) \Big\} \\
&= \pi(\boldsymbol{\beta}) \exp\{ -\frac{n}{2}(\boldsymbol{\beta} - \hat{\boldsymbol{\beta}})' I_n (\boldsymbol{\beta} - \hat{\boldsymbol{\beta}}) + Q_n \},
\end{aligned} \tag{.5}$$

where, by (.4),

$$\begin{aligned}
Q_n &= -\frac{n}{2}(\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}_0)' V_{12}' V_{11}^{-1} V_{12} (\boldsymbol{\beta} - 2\hat{\boldsymbol{\beta}} + \boldsymbol{\beta}_0) + \frac{n}{2}(\boldsymbol{\beta} - \hat{\boldsymbol{\beta}})' V_{12}' V_{11}^{-1} V_{12} (\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}_0) \\
&\qquad - \frac{1}{2} W_n' V_{11}^{-1} W_n + o_P(1) \\
&= \frac{n}{2}(\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}_0)' V_{12}' V_{11}^{-1} V_{12} (\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}_0) - \frac{1}{2} W_n' V_{11}^{-1} W_n + o_P(1) \\
&= \frac{n}{2}(n^{-1/2}(V_{12}' V_{11}^{-1} V_{12})^{-1} V_{12}' V_{11}^{-1} W_n + o_P(n^{-1/2}))' V_{12}' V_{11}^{-1} V_{12} \\
&\qquad \times (n^{-1/2}(V_{12}' V_{11}^{-1} V_{12})^{-1} V_{12}' V_{11}^{-1} W_n + o_P(n^{-1/2})) - \frac{1}{2} W_n' V_{11}^{-1} W_n + o_P(1) \\
&= \frac{1}{2} W_n' V_{11}^{-1} W_n + o_P(1) - \frac{1}{2} W_n' V_{11}^{-1} W_n + o_P(1) = o_P(1).
\end{aligned} \tag{.6}$$

Therefore (2.16) follows from (.5) and (.6) together with $\log(\pi(\boldsymbol{\beta})) = \log(\pi(\boldsymbol{\beta}_0)) + O(n^{-1/2})$ for $\boldsymbol{\beta} - \boldsymbol{\beta}_0 = O(n^{-1/2})$ owing to Assumption 6.

The remaining part of Theorem 2.3 can be proved, by using Assumption 6, as done in the corresponding proof of Theorem 3.2 of Lu et al. (2007, pp. 186). The details are therefore omitted.

REFERENCES

Berlinet, A., Vajda, I., and Van der Meulen, E. (1998). About the asymptotic accuracy of barron density estimates. *Information Theory, IEEE Transactions*, 44(3):999–1009.

Bickel, P. and Fan, J. (1996). Some problems on the estimation of unimodal densities. *Statistica Sinica*, 6:23–46.

Birgé, L. (1997). Estimation of unimodal densities without smoothness assumptions. *The Annals of Statistics*, 25(3):970–981.

Brunner, L. (1992). Bayesian nonparametric methods for data from a unimodal density. *Statistics & Probability letters*, 14(3):195–199.

Chernozhukov, V. and Hong, H. (2003). An mcmc approach to classical estimation. *Journal of Econometrics*, 115(2):293–346.

Dunson, D., Pillai, N., and Park, J. (2007). Bayesian density regression. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 69(2):163–183.

Eddy, W. (1980). Optimum kernel estimators of the mode. *The Annals of Statistics*, 8(4):870–882.

Feller, W. (1971). *An introduction to probability theory and its applications*, volume 2. Wiley-New york.

Gasser, T., Hall, P., and Presnell, B. (1998). Nonparametric estimation of the mode of a distribution of random curves. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 60(4):681–691.

Grenander, U. (1965). Some direct estimates of the mode. *The Annals of Mathematical Statistics*, 36(1):131–138.

Hall, P. and Huang, L. (2001). Nonparametric kernel regression subject to monotonicity constraints. *The Annals of Statistics*, 29(3):624–647.

Hall, P., Peng, L., and Rau, C. (2001). Local likelihood tracking of fault lines and boundaries. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 63(3):569–582.

Heckman, D., Geiser, D., Eidell, B., Stauffer, R., Kardos, N., and Hedges, S. (2001). Molecular evidence for the early colonization of land by fungi and plants. *Science*, 293(5532):1129.

Hedges, S. and Shah, P. (2003). Comparison of mode estimation methods and application in molecular clock analysis. *BMC Bioinformatics*, 4(1):31.

Kemp, G. C. and Santos Silva, J. (2012). Regression towards the mode. *Journal of Econometrics*, 170(1):92–101.

Kumar, S. and Hedges, S. (1998). A molecular timescale for vertebrate evolution. *Nature*, 392(6679):917–920.

Lee, M. (1989). Mode regression. *Journal of Econometrics*, 42(3):337–349.

Lee, M. (1993). Quadratic mode regression. *Journal of Econometrics*, 57(1-3):1–19.

Lu, Z., Tjstheim, D., and Yao, Q. (2007). Adaptive varying-coefficient linear models for stochastic processes: asymptotic theory. *Statistica Sinica*, 17:177–197.

Manski, C. (1991). Regression. *Journal of Economic Literature*, 29(1):34–50.

Markov, H., Valtchev, T., Borissova, J., and Golev, V. (1997). An algorithm to "clean" close stellar companions. *Astronomy and Astrophysics Supplement Series*, 122(1):193–199.

Meyer, M. (2001). An alternative unimodal density estimator with a consistent estimate of the mode. *Statistica Sinica*, 11(4):1159–1174.

Molanes Lopez, E., Keilegom, I., and Veraverbeke, N. (2009). Empirical likelihood for non-smooth criterion functions. *Scandinavian Journal of Statistics*, 36(3):413–432.

Parzen, E. (1962). On estimation of a probability density function and mode. *The Annals of Mathematical Statistics*, 33(3):1065–1076.

Qin, J. and Lawless, J. (1994). Empirical likelihood and general estimating equations. *The Annals of Statistics*, pages 300–325.

Roberts, G. and Rosenthal, J. (2001). Optimal scaling for various metropolis-hastings algorithms. *Statistical Science*, 16(4):351–367.

Silverman, B. (1986). *Density estimation for statistics and data analysis*, volume 26. Chapman & Hall/CRC.

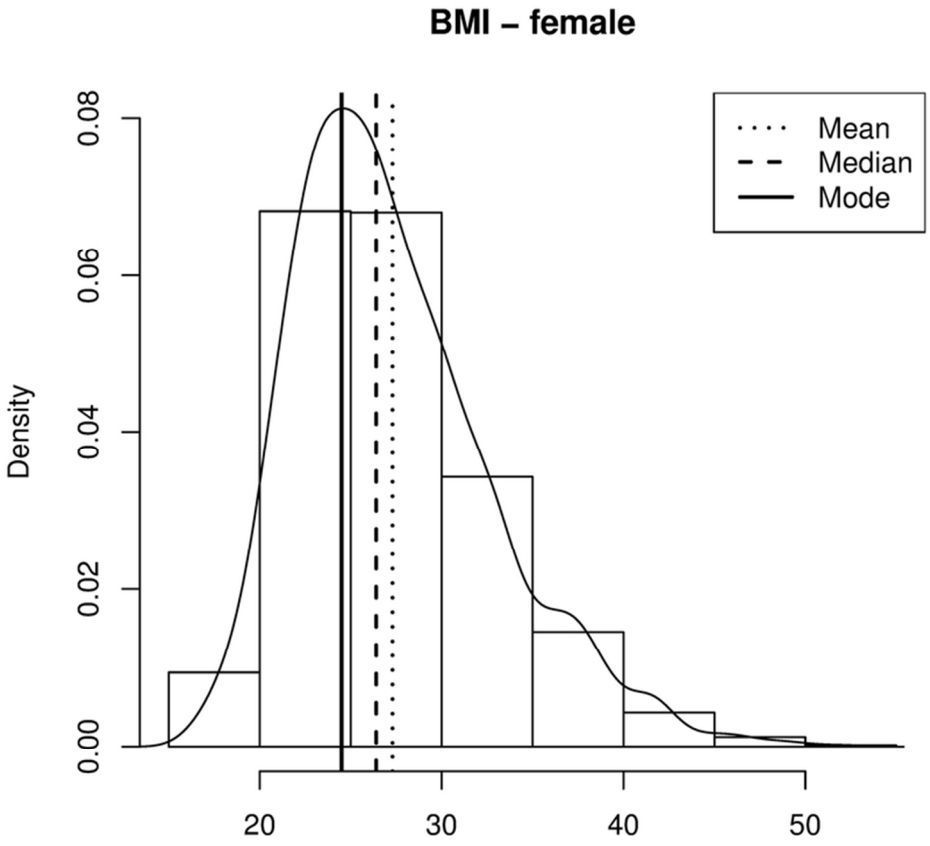
Smith, B. (2007). boa: an r package for mcmc output convergence assessment and posterior inference. *Journal of Statistical Software*, 21(11):1–37.

Van der Vaart, A. (1998). *Asymptotic statistics*. Cambridge Univ Pr.

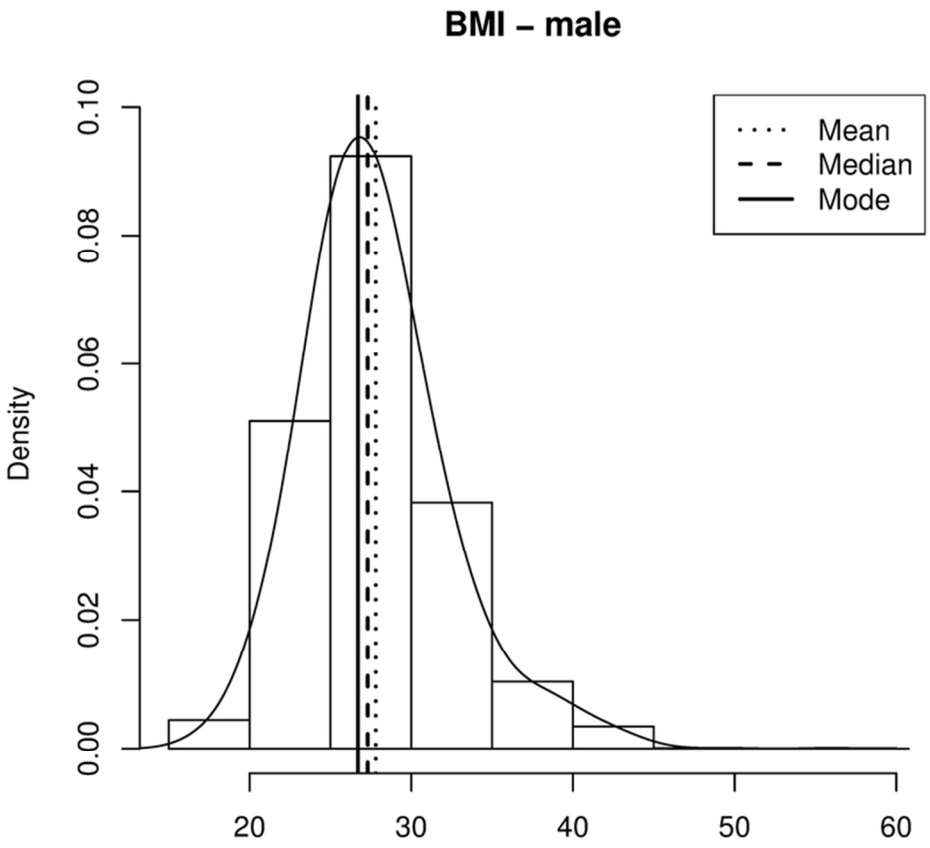
Yang, Y. and He, X. (2012). Bayesian empirical likelihood for quantile regression. *Annals of Statistics*, 40(12):1102–1131.

Yao, W. and Li, L. (2013). A new regression model: modal linear regression. *Scandinavian Journal of Statistics*.

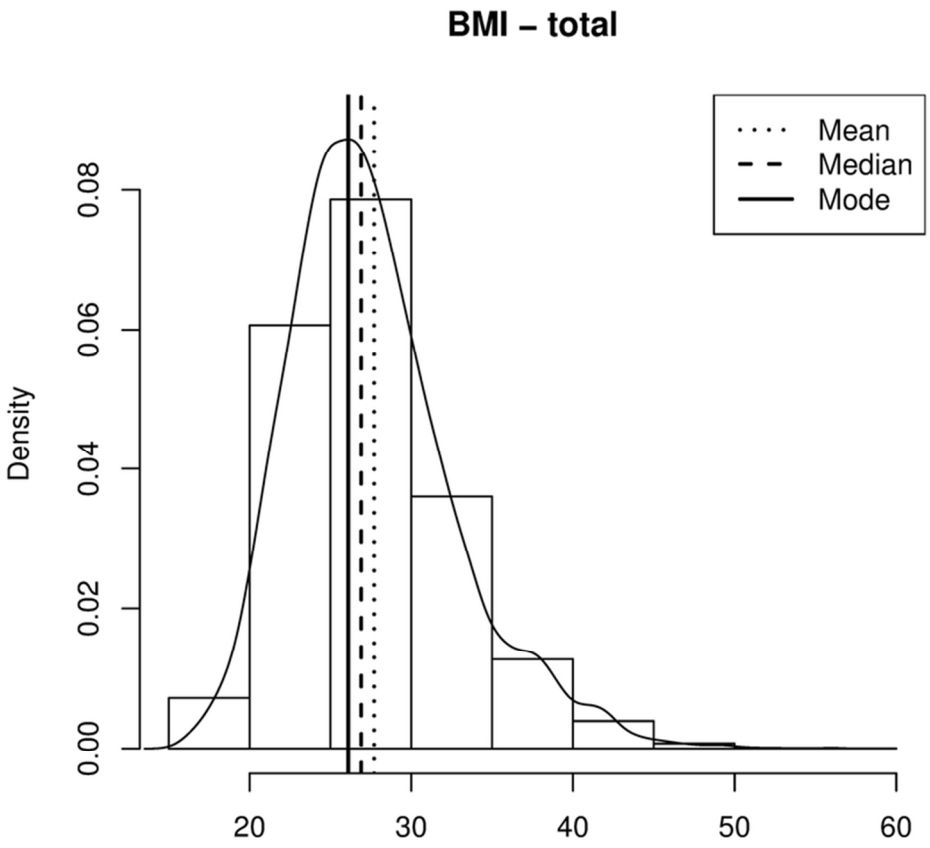
Yasukawa, K. (1926). On the probable error of the mode of skew frequency distributions. *Biometrika*, 18(3/4):263–292.



145x145mm (150 x 150 DPI)



145x145mm (150 x 150 DPI)



145x145mm (150 x 150 DPI)