Convolution Mode Regression

Eduardo Schirmer Finn*§, Eduardo Horta*†

Abstract

For highly skewed or fat-tailed distributions, mean or median-based methods may be inadequate to capture centrality in the data. This deficiency of traditional methods has fostered the emergence of conditional mode models as a valuable approach. However, estimating the conditional mode of a variable given certain covariates presents challenges: nonparametric approaches suffer from the "curse of dimensionality", while the semiparametric strategy can lead to non-convex optimization problems. We propose a novel estimator for mode regression, constructed by inverting the conditional quantile density. Unlike existing approaches in the literature, we estimate the quantile density function by inverting a convolution-type smoothed variant of the quantile regression model. Our resulting estimator is consistent, with the benefit of having uniform convergence with respect to both the design points of the covariates and to the bandwidth.

Keywords: *Mode Regression; Convolution-based Smoothing; Conditional Quantile; Asymptotic Theory; Uniform Convergence.*

1 Introduction

Conventional econometric methods are generally mean-based; such methods may fail to express the central tendency if distributions are highly skewed or long-tailed (Kemp and Santos-Silva, 2012; Chen et al., 2016). The conditional mode emerges as a robust alternative, conveying the desirable interpretation of being the most likely value of a dataset (Chacón, 2020). This interpretation becomes particularly valuable when dealing with continuous variables, which, unlike discrete random variables, do not have a straightforward sample mode version. In such cases, the mode is formally defined as a point of maximum (local or global) for the conditional probability density function. Since Lee (1989), the estimation of conditional mode, called mode regression, has demonstrated its utility across various domains, specially in applications with asymmetric data, such as wages (Zhang, Kato, and Ruppert, 2023); electrical energy consumption (Ota, Kato, and Hara, 2019); medical sciences (Wang et al., 2017), traffic data (Einbeck and Tutz, 2006) and a forest fire dataset (Yao and Li, 2014).

In reviewing the conditional mode literature, two fundamental considerations emerge in the discussion: firstly, the assumption of whether the mode is global or local; secondly,

*Department of Statistics, Federal University of Rio Grande do Sul (UFRGS), Porto Alegre, Brazil.
§Corresponding Author, eduardosfinn@outlook.com
†eduardo.horta@ufrgs.br

if the estimation strategy employed is semiparametric (linear) or fully nonparametric. Linear approaches require the condition that the mode is a unique maximizer of the conditional density; nonparametric techniques in turn are usually used for multimodal models, but not exclusively—in a few cases the global mode is also estimated nonparametrically (Sager and Thisted, 1982; Feng et al., 2020). The first semiparametric estimator considering a unique mode was developed by Lee (1989) and establishes a linear relationship between the mode of the response and the covariates; however, despite being elegant, this model is impractical;[1] subsequent models, such as the ones of Kemp and Santos-Silva (2012) and Yao and Li (2014), yield out non-convex optimization problems, resulting in functions that may have multiple maxima; also, the algorithms developed are sensible to the starting points. On the other hand, nonparametric estimation tends to avoid misspecification (Yao et al., 2012; Chen et al., 2016); still, these methods suffer from the "curse of dimensionality" and, moreover, have slow convergence rates (Zhang, Kato, and Ruppert, 2023). Two related methods have been developed by Ota, Kato, and Hara (2019) and Zhang, Kato, and Ruppert (2023); both relying on estimating the conditional quantile density by inverting a quantile regression, overcoming slow convergence and optimization issues.

Similarly to what is done in Ota, Kato, and Hara (2019) and Zhang, Kato, and Ruppert (2023), we propose a novel approach which relies on working with a different quantile regression estimator, the smoothed version of Fernandes, Guerre, and Horta (2021). This alternative of the estimator provides some advantages: it is continuous, asymptotically unbiased and less variable than the traditional quantile regression estimator. Previous simulations (Ongaratto and Horta, 2021) have shown that our estimator outperforms the Ota, Kato, and Hara's (2019) in many aspects.

The main goal of this work is to derive asymptotic properties for the estimator of the conditional mode via smoothed quantile regression, hereby denominated *Convolution Mode Regression*, proving its consistency, with convergence rates. Additionally, our approach differs from Zhang, Kato, and Ruppert (2023) as they opt for an *"estimate then smooth"* procedure (using the traditional quantile regression framework), while we *"smooth then estimate"* (a smoothed and continuous version of the estimator). The "estimate then smooth" approach is more akin to the proposition of Parzen (1979), of quantile estimation by smoothing the sample quantile function; whereas the "smooth then estimate" approach parallels Nadaraya (1964)'s method, in which the unconditional quantile is estimated via inverting a smoothed estimator of the cdf (cumulative distribution function). The convergence rate of the Zhang, Kato, and Ruppert (2023) estimator is $O_P(n^{-1/2}h^{-3/2}\sqrt{\log n}+h^2)$, whilst ours is $O_P((\frac{\log n}{nh})^{1/4}) + o(h^{1/2})$; both rates free from the "curse of dimensionality". Nonetheless, our estimator can be more adequate in cases where the choice of the bandwidth is data-driven, due to its uniform convergence also in h, something not present in

[1] According to Kemp and Santos-Silva (2012) this model has restrictive assumptions on the conditional density of the response and, due to the objective function, the estimator lacks a tractable distribution.

Zhang, Kato, and Ruppert (2023).

1.1 Literature Review

Estimation of a global mode in a continuous variable environment is not as explicit as it seems. Sager (1978) and Chacón (2020) divide the mode estimators into two categories, direct and indirect. The latter classification is for when an intermediate step is required, such as estimating the density function (as it is not known) which is typically done via Kernel Density Estimation (KDE), as firstly presented by Parzen (1962). Estimators are of the direct kind when they are specifically constructed for the sole purpose of estimating the mode, this is the case for the "naive" estimator of the mode based on Chernoff (1964). As stated by Sager (1978), these classifications may blur, as the majority of direct estimators reveal some kind of linkage with a type of density estimation (Chacón, 2020).

Following the introduction of mode estimators, attention shifted towards studying how the mode of a variable of interest responds to covariates. Sager and Thisted (1982) generalized the framework of Chernoff (1964) and developed the first mode regression. The initial model established that the global mode of the dependent variable is a monotone function of the covariate, and was estimated via a maximum likelihood nonparametric estimator. Despite this model's limitation in being applicable only to ordinal data, it laid the groundwork for mode regression. Additionally, it was shown that the conditional mode estimator could be formulated by applying a plug-in from a density estimator, such as KDE; nevertheless, consistency was not achieved. Remarkably, such combination of nonparametric estimation alongside the assumption of a unique (global) mode, is not very common in the literature.

Building on this initial work, mode regression has evolved and can be categorized according to two major factors: (i) unimodal vs. multimodal assumption, and (ii) semiparametric vs. nonparametric estimation. Firstly, (i) the assumption of the mode being unique, referred to as *global mode*; or assuming more than one point of maxima and that the data distribution is multimodal, namely *local mode* assumption. Secondly, (ii) the type of estimation that is employed, which can be done semiparametrically, mainly formulating a linear relationship between the dependent variable and the covariates; vis-à-vis estimating the conditional mode nonparametrically in order to allow for multiple local modes, as in Chen et al. (2016). It is important to note that while all linear/semiparametric mode regressions require the assumption of a global mode, the inverse statement does not hold, that is, there is no need for a model with a unique mode to be linear, since it can be estimated nonparametrically. On the other hand, our review did not uncover a multimodal model that was not estimated nonparametrically. In light of this, we can divide the conditional mode estimation literature into 4 different strands[2]: (1) unique mode with linear/ semiparametric estimation; (2) unique mode with nonparametric estimation,

[2]Some papers may not fit precisely in this categorization, since they mix parametric and nonparametric traits (Liu et al., 2013; Wang, 2024), or use a Bayesian approach (Yu and Aristodemou, 2012).

and (3) multimodal with nonparametric estimation; afterwards, special attention is given to (4) conditional quantile approaches towards the mode, since this is more related to our contribution. The first and last literature strands are the ones that will be further discussed in this article, since they bear relevance to the model we propose. For a general survey of the role of the mode in statistics, we recommend Chacón (2020); especially for mode regression, we indicate the review of Chen (2018).

(1) Unique Mode & Semiparametric Estimation: following the initial work of Sager and Thisted (1982), Lee (1989) proposed a linear approach where a smoothed loss function is used. The main drawback of this line of action lies in the underlying assumptions, specially homogeneity and symmetry of the error terms, which led to the conditional mean coinciding with the conditional mode. Also, the estimator is impractical due to its distribution being intractable because of the objective function. This work inspired further investigation regarding linear conditional mode estimation, such as Lee (1993), where the rectangular kernel was replaced for a quadratic one. Still, restrictive assumptions were required on the conditional density of the response variable (Chen, 2018). Lee's papers inspired further work that sought to eliminate the assumption of a symmetric error term, such as: Kemp and Santos-Silva (2012), where estimation is done via minimization of a kernel-based loss function; as well as Yao and Li (2014), who focused on high-dimensional data. Both methods have algorithmic issues, leading to a nonconvex optimization problems, with no guarantee of convergence to the global maximum, and high sensitivity to the selected starting point. Some further exploration of this literature strand can be found in variable selection (Zhang et al., 2013), time series analysis (Kemp et al., 2020) and in panel data (Ullah et al., 2021).

(2) Unique Mode & Nonparametric Estimation: a linear approach can be too restrictive depending on the type of data; thus, nonparametric regression can model the components of the conditional mode as smooth functions of the covariates (Chen et al., 2016). Apart from the pioneering work of Sager and Thisted (1982), the global mode is estimated nonparametrically in Yao et al. (2012) by applying local polynomial smoothing. In the generalization provided by the authors, when the degree of the polynomial is zero and there is a single covariate, the method is referred to as local linear modal regression; if there are no covariates, this method reduces to a kernel density estimate. Nonetheless, the model of Yao et al. (2012) carries out some issues: the application is limited to unique mode regression (Chen, 2018), it suffers from the "curse of dimensionality" and symmetry for the error's distributions is imposed.[3] A more recent nonparametric regression for single mode that mitigates the "curse of dimensionality" is found in Feng et al. (2020), where a statistical learning analysis is applied to mode regression. The estimation of the conditional mode is achieved via an empirical risk minimization approach. Such modulation turns the problem into non-dependable on dimension, thus being applicable

[3]According to Zhang, Kato, and Ruppert (2023), their sixth assumption (symmetry of the error term) leads to the problem corresponding to conditional mean estimation.

to areas such as big data and machine learning.

(3) Multimodal & Nonparametric Estimation: as it is not always the case that data structures can be interpreted as unimodal, multimode regressions come forth as alternatives that enable to uncover hidden relations otherwise undetected (Chen, 2018). The proposal to consider various modes arises from Scott (1992), who defined them as points of local maxima of the conditional density of the response. Such is the case for Matzner-Løfber et al. (1998), where a forecasting comparison is carried out for three kernel-based methods, namely, conditional mean regression, conditional median regression and conditional mode regression. The findings indicate that, when dealing with bimodal data, mode regression outperforms the competitors in terms of accuracy. Motivated by a similar prediction problem, where the data has two pronounced modes, Einbeck and Tutz (2006) develop the first systematic investigation regarding local modes (Chen, 2018; Chacón, 2020). Their estimator is computed from a modified *meanshift* algorithm; subsequently, it is applied to traffic data, and is used to determine the modal speed at different flows of cars. In contrast to Einbeck and Tutz (2006), where there is a lack of asymptotic theory, Chen et al. (2016) develop a conditional (multi)mode nonparametric model based on KDE. Less restrictive assumptions on the kernel density function are used, the method yields strong asymptotic properties and model misspecification is avoided. Despite this, according to Zhang, Kato, and Ruppert (2023) the convergence rate of the estimator is slow even when the number of covariates is not too large, namely, the approach suffers from the "curse of dimensionality".

(4) Conditional Quantile Approach: motivated by the fact that linear mode regression models resulted in nonconvex optimization problems, and also in possible model misspecification, whereas the flexibility from nonparametric estimators comes at the cost of the "curse of dimensionality", Ota, Kato, and Hara (2019) developed a novel semiparametric approach based on quantile regression. The main idea is to use this regression framework as an intermediate step for conditional mode estimation: namely, the quantile function of Koenker and Bassett (1978) is estimated, from which a conditional *quantile density* estimator is obtained via numerical differentiation. Importantly, the underlying model does not impose linearity of the mode function, not even when the quantile regression model used is linear-in-covariates. Furthermore, this approach avoids the "curse of dimensionality" and is appealing computationally, since the quantile regression estimator can be written as a linear programming problem. Still, using the traditional estimator for the quantile function can bring some concerns, since the empirical conditional quantile function has jumps—hence the mentioned numerical differentiation. In order to surpass this problem, Zhang, Kato, and Ruppert (2023) propose to post-smooth the quantile regression estimator by a kernel function. Not only does this strategy circumvents numerical differentiation, but also it yields faster convergence rates and an estimator that is asymptotically Normal, in contrast to a nonstandard Chernoff distribution as in Ota, Kato, and Hara (2019). Both models take off from a key identity that we also explore in this paper, namely, that the quantile density is the reciprocal of the density function, evaluated at the quantile of interest, which summarizes how the conditional mode can be retrieved from the quantile density.

1.2 Organization

The rest of this paper is organized as follows. In Section 2 the model is presented, we introduce the estimator and explore its relationship with the smoothed quantile regression. In Section 3 we enunciate the main mathematical results of the paper, as well as the needed assumptions for them to hold; also, we compare our convergence rates to the most similar model in the literature for some different bandwidth scenarios. In Section 4 we state our concluding remarks along with possibilities for future work. The Appendix contains additional mathematical material, such as derivations and convergence rates calculations.

2 Convolution Mode Regression

2.1 Setup

Let $Y \in \mathbb{R}$ represent a target random variable for which we are interested in estimating the conditional mode, given a d-dimensional vector X of covariates, and write $\mathcal{X} := \text{support}(X)$. Assume that $Y|X = x$ is continuous and unimodal, having conditional cdf $F(\cdot|x)$ and conditional pdf $f(\cdot|x)$. Then, the **conditional mode** of Y given $X = x$, denoted by $m(x)$, is defined as:

$$m(x) = \arg\max_{y \in \mathbb{R}} f(y|x), \qquad x \in \mathcal{X}. \tag{1}$$

Thus, $m(x)$ corresponds to the point in the covariate space at which the (conditional) density of the response attains its maximum value. Additionally, define the **τ-th conditional quantile of Y given** $X = x$ as the scalar $Q(\tau|x)$ given by

$$Q(\tau|x) := \inf\{y \in \mathbb{R} : F(y|x) \geqslant \tau\}, \quad \tau \in (0,1), x \in \mathcal{X}$$

and the **conditional quantile function** as the mapping $\tau \mapsto Q(\tau|x)$.

It is important to point out that the quantile function is entirely retrievable from the cdf, since it is just the generalized inverse of the function $y \mapsto F(y|x)$, and in the case of the distribution being a continuous function we have that $Q(\cdot|x) := F^{-1}(\cdot|x)$ for each x (van der Vaart, 1998; Koenker, 2005). Furthermore, the **conditional quantile density** is defined through

$$q(\tau|x) = Q'(\tau|x) = \frac{\partial Q(\tau|x)}{\partial \tau}, \quad \tau \in (0,1),\ x \in \mathcal{X}. \tag{2}$$

A key identity explored by Ota, Kato, and Hara (2019) and Zhang, Kato, and Ruppert

(2023) is that, as a consequence of the Inverse Function Theorem, the identity

$$q(\tau|x) = \frac{1}{f(Q(\tau|x)|x)} \tag{3}$$

holds for every allowable x and τ. In this sense, given some regularity conditions which we introduce below, we can minimize the inverse of the density, as in equation (3), and retrieve the maximizer of $y \mapsto f(y|x)$ from equation (1); thus,

$$m(x) = Q\big(\arg\min_\tau q(\tau|x) \,|\, x\big) \tag{4}$$

Regarding the quantile function, both Ota, Kato, and Hara (2019) and Zhang, Kato, and Ruppert (2023) consider the quantile regression model developed by Koenker and Bassett (1978), which stipulates a linear-in-covariates representation of the conditional quantile function:

$$Q(\tau|x) = x^\intercal \beta(\tau), \quad \tau \in (0,1),\, x \in \mathcal{X}, \tag{5}$$

where $\beta : (0,1) \mapsto \mathbb{R}^d$ is a functional parameter. For each fixed τ in the interval $(0,1)$, the vector $\beta(\tau)$ in (5) solves a similar minimization problem as the one found in classic linear regression. For this end, the following population objective function is proposed:

$$R(b;\tau) := \mathbb{E}[\rho_\tau(Y - X^\intercal b)] = \int \rho_\tau(t)\,\mathrm{d}F(t;b) \tag{6}$$

with $\rho_\tau(u) := u[\tau - \mathbb{I}(u < 0)]$ known as the check function. The true parameter $\beta(\tau)$ minimizes $R(b;t)$ with respect to $b \in \mathbb{R}^d$. The sample equivalent proposed by Koenker and Bassett (1978) is defined as:

$$\widehat{R}(b;\tau) := \frac{1}{n}\sum_{i=1}^n \rho_\tau(Y_i - X_i^\intercal b) = \int \rho_\tau(t)\,\mathrm{d}\widehat{F}(t;b), \tag{7}$$

where $\widehat{F}(\cdot;b)$ is the empirical distribution function of $\varepsilon_i(b) := Y_i - X_i^\intercal b$, for $i = 1, ..., n$, with the traditional quantile regression estimator as the minimizer of $\widehat{R}(b;\tau)$, with respect to $b \in \mathbb{R}^d$, that is:

$$\widehat{\beta}(\tau) = \arg\min_{b\in\mathbb{R}^d} \widehat{R}(b;\tau), \qquad \tau \in (0,1) \tag{8}$$

According to Theorem 2.1 in Bassett and Koenker (1982), the empirical conditional quantile function $\tau \mapsto x^\intercal\widehat{\beta}(\tau)$ exhibits jumps, in particular it is not differentiable. To overcome this issue, Fernandes, Guerre, and Horta (2021) proposed using a kernel-type cdf estimator, similar to Nadaraya (1964), instead of the empirical distribution function. The resulting smoothed version of the sample objective function in (7) is:

$$\widehat{R}_h(b;\tau) = \frac{1}{n}\sum_{i=1}^n k_h * \rho_\tau(Y_i - X_i^\intercal b) = \int \rho_\tau(t)\widehat{f}_h(t;b)\,\mathrm{d}t \tag{9}$$

where the symbol $*$ denotes the convolution operator, and where $\widehat{f}_h(\cdot; b)$ is the kernel estimator of the density of $Y_i - X_i^\intercal b$. Here, $k_h(u) = k(u/h)/h$, where $k\colon \mathbb{R} \to \mathbb{R}_+$ is a smooth kernel function and $h > 0$. The new estimator is the minimizer of the objective function (9), called the **smoothed quantile regression estimator (SQRE)** and defined by:

$$\widehat{\beta}_h(\tau) := \arg\min_{b\in\mathbb{R}^d} \widehat{R}_h(b;\tau), \qquad \tau \in (0,1). \tag{10}$$

The mapping $\tau \mapsto \widehat{\beta}_h(\tau)$ is continuously differentiable over the interval $(0, 1)$, unlike $\widehat{\beta}$. Differentiability offers notable advantages, and the reasons are twofold: (i) the smoothness of the objective function ensures the regularity of the resulting estimator; (ii) the asymptotic covariance matrix of $\widehat{\beta}_h(\tau)$ can be estimated in a standard fashion, as in Newey and McFadden (1994). Regarding differentiability, writing $\widehat{R}_h^{(1)}(b; \tau) := \partial\widehat{R}_h(b; \tau)/\partial b$, the SQRE satisfies the first-order condition $\widehat{R}_h^{(1)}(\widehat{\beta}_h(\tau); \tau) = 0$. Accordingly, following the Implicit Function Theorem, we obtain:

$$\widehat{\beta}_h^{(1)}(\tau) = \frac{\partial\widehat{\beta}_h(\tau)}{\partial\tau} := \left[\widehat{R}_h^{(2)}\left(\widehat{\beta}_h(\tau);\tau\right)\right]^{-1}\bar{X} \tag{11}$$

Explicit formulas for the first and second order derivatives of $\widehat{R}_h(b; \tau)$ with respect to b (respectively, $\widehat{R}_h^{(1)}(b; \tau)$ and $\widehat{R}_h^{(2)}(b; \tau)$) are provided in equation (33) in A.1.

2.2 Estimation

Consider the following objective (or "sparsity"[4]) function:

$$s_x(\tau) := -q(\tau|x), \quad \tau \in (0,1),\, x \in \mathcal{X}. \tag{12}$$

It is not difficult to show that

$$s_x(\tau) = -x^\intercal \beta^{(1)}(\tau) = -x^\intercal[D(\tau)]^{-1}\mathbb{E}X \tag{13}$$

where

$$D(\tau) = R^{(2)}(\beta(\tau);\tau) = \mathbb{E}[XX^\intercal f(X^\intercal\beta(\tau)|X)]. \tag{14}$$

Under some regularity assumptions that will be introduced below, the function $\tau \mapsto s_x(\tau)$ has a unique maximizer, denoted τ_x, which we call the **conditional quantile mode** of Y given $X = x$. If we plug in this optimizer in the quantile function $Q(\cdot|x)$ we get the expression $m(x) = Q(\tau_x|x)$. Consequently, the estimation of $m(x)$ boils down to estimating the conditional quantile function, and τ_x. In view of (13), we define, for conformable τ, x and h, the **sample conditional sparsity function** as:

$$\widehat{s}_{x,h}(\tau) = -x^\intercal\widehat{\beta}_h^{(1)}(\tau) = -x^\intercal[\widehat{D}_h(\tau)]^{-1}\bar{X}, \tag{15}$$

[4] This is the nomenclature used by Ota, Kato, and Hara (2019) and Zhang, Kato, and Ruppert (2023).

where

$$\widehat{D}_h(\tau) = \widehat{R}_h^{(2)}(\widehat{\beta}_h(\tau);\tau) = \frac{1}{n}\sum_{i=1}^{n} X_i X_i^{\intercal} k_h(X_i^{\intercal}\widehat{\beta}_h(\tau) - Y_i), \tag{16}$$

see Fernandes, Guerre, and Horta (2021).

The optimizers for the sparsity functions, both population and sample, as defined in (13) and (15), are given by:

$$\tau_x = \arg\max_{\tau\in(0,1)} s_x(\tau) \quad \text{and} \quad \widehat{\tau}_{x,h} = \arg\max_{\alpha\leqslant\tau\leqslant 1-\alpha} \widehat{s}_{x,h}(\tau) \tag{17}$$

where $0 < \alpha < 1/2$ is a constant.

Our proposed **smoothed conditional mode estimator** is then given by

$$\widehat{m}_h(x) := \widehat{Q}_{x,h}(\widehat{\tau}_{x,h}) = x^{\intercal}\widehat{\beta}_h(\widehat{\tau}_{x,h}), \tag{18}$$

for all $x \in \mathcal{X}$ and every allowable h.

3 Main Results

Before providing consistency results of the proposed estimator, we state the conditions for which our results are derived.

3.1 Assumptions:

- **A1:** The support of X, denoted $\mathcal{X}$, is compact and a subset of $\bar{\mathbb{R}}^d_{+*}$, i.e., the components of X are positive, bounded RVs. The matrix $\mathbb{E}[XX^{\intercal}]$ is full rank.
- **A2:** The mapping $\tau \mapsto \beta(\tau)$ is three times continuously differentiable.
- **A3:** The conditional density $f(y|x)$ is continuous and strictly positive over $\mathbb{R}\times\mathcal{X}$. Also, the derivative $f^{(1)}(\cdot|\cdot)$ exists and is uniformly continuous in the sense that

  $$\lim_{\epsilon\to 0} \sup_{(x,y)\in\mathbb{R}^{d+1}} \sup_{t:|t|\leqslant\epsilon} \left| f^{(1)}(y+t|x) - f^{(1)}(y|x)\right| = 0,$$

  and that $\sup_{(x,y)\in\mathbb{R}^{d+1}} |f^{(j)}(y|x)| < \infty$ and $\lim_{y\to\pm\infty} f^{(j)}(y|x) = 0$ for all $j \in \{0,1\}$.

  Remark. The degree of differentiability of $f(\cdot|\cdot)$ is used in Fernandes, Guerre, and Horta (2021) to control the order of the smoothing kernel. Here, we set the maximum value of j equal to 1, for simplicity.

- **A4:** The kernel $k\colon \mathbb{R} \mapsto \mathbb{R}$ is even, integrable and has bounded first and second derivatives. Additionally, $\int k(z)\mathrm{d}z = 1$; $0 < \int_0^\infty K(z)[1-K(z)]\mathrm{d}z < \infty$ and, lastly, $0 < \int z^2 k(z)\mathrm{d}z < \infty$.
- **A5:** $h \in [\underline{h}_n, \bar{h}_n]$ with $n\underline{h}_n^3/\log n \to \infty$ and $\bar{h}_n = o(1)$.

- **A6:** For all $x \in \mathcal{X}$, there exists $\tau_x \in (0,1)$ such that, for every $\epsilon > 0$, it holds that

$$\sup_{\tau:|\tau-\tau_x|\geqslant\epsilon} s_x(\tau) < s_x(\tau_x)$$

- **A7:** For some $0 < \alpha < 1/2$, it holds that

$$\alpha < \inf_{x\in\mathcal{X}} \tau_x \leqslant \sup_{x\in\mathcal{X}} \tau_x < 1-\alpha.$$

Assumptions A1-A5 are taken directly from Fernandes, Guerre, and Horta (2021), with minor modifications. Due to A1-A3, the Hessian $D(\tau)$ as defined in (14), is positive definite for all possible values of $\tau \in (0,1)$, therefore, $D(\tau)$ is invertible. Additionally, A2 ensures the function $\tau \mapsto Q(\tau|x)$ is increasing over the interval $(0,1)$, and, together with A3, that its derivative with respect to τ is strictly positive. Also, A3 expresses some ordinary regularity conditions which guarantee smoothness of $f(\cdot|\cdot)$ (Koenker, 2005). Similar conditions can be found in Chen et al. (2016); Ota, Kato, and Hara (2019); Zhang, Kato, and Ruppert (2023); however, each of these estimates requires four-times continuous differentiability of the density. Assumptions A4 and A5 concern the kernel function k and the bandwidth parameter h. A6 ensures uniqueness of the conditional mode and is also commonly used in deriving consistency of M-estimators, see Theorem 5.7 in van der Vaart (1998). Finally, Assumption A7 limits the possible values for the optimizer τ_x, ensuring that the conditional modes are bounded away from the tails of the conditional distributions, uniformly on the covariate space.

3.2 Consistency of the Convolution Mode Regression Estimator:

The following lemma is a reinstatement of an inequality in Fernandes, Guerre, and Horta (2021).

Lemma 1. Under Assumption A1 to A5, it holds that

$$\left\|\widehat{D}_h(\tau) - D(\tau)\right\| = o(h^j) + O_P\left(\sqrt{\log n/(nh)}\right), \tag{19}$$

uniformly for $\tau \in [\alpha, 1-\alpha]$ and $h \in [\underline{h}_n, \bar{h}_n]$.

Proof. See the proof of Proposition 1 in Fernandes, Guerre, and Horta (2021). ■

Our next result is regarding the sample sparsity function and the fact that it converges to the population counterpart.

Lemma 2. Under Assumptions A1 to A5, it holds that

$$|\widehat{s}_{x,h}(\tau) - s_x(\tau)| = o(h) + O_P\left(\sqrt{\log n/(nh)}\right)$$

uniformly over $\tau \in [\alpha, 1-\alpha]$, $x \in \mathcal{X}$ and $h \in [\underline{h}_n, \bar{h}_n]$.

Proof. Write

$$-(\widehat{s}_{x,h}(\tau) - s_x(\tau)) \;=\; x^\intercal [\widehat{D}_h(\tau)]^{-1} \bar{X} \;-\; x^\intercal [D(\tau)]^{-1} \mathbb{E}X.$$

Using $\bar{X} = \bar{X} - \mathbb{E}X + \mathbb{E}X$ and rearranging, we have

$$\begin{aligned} -(\widehat{s}_{x,h}(\tau) - s_x(\tau)) &= x^\intercal [\widehat{D}_h(\tau)]^{-1} (\bar{X} - \mathbb{E}X) \\ &\qquad + \; x^\intercal \left\{ [\widehat{D}_h(\tau)]^{-1} - [D(\tau)]^{-1} \right\} (\mathbb{E}X) \end{aligned} \tag{20}$$

Lemma 1 implies, by the local Lipschitz property of matrix inversion, that

$$\left\| [\widehat{D}_h(\tau)]^{-1} - [D(\tau)]^{-1} \right\| = o(h) + O_P\left(\sqrt{\log n/(nh)}\right)$$

uniformly in τ and h as above. This together with

$$\sup_{\tau,h} |\widehat{D}_h(\tau)| = O_P(1), \quad \bar{X} - \mathbb{E}X = O_P(1/\sqrt{n}), \quad \mathbb{E}X = O(1), \quad \sup_{x \in \mathcal{X}} \|x\| = O(1)$$

tells us that

$$\begin{aligned} -(\widehat{s}_{x,h}(\tau) - s_x(\tau)) &= x^\intercal O_p(1) O_p(1/\sqrt{n}) \\ &\qquad + x^\intercal \left(o(h) + O_P\left(\sqrt{\log n/(nh)}\right)\right) O(1) \\ &= o(h) + O_P\left(\sqrt{\log n/(nh)}\right) \end{aligned} \tag{21}$$

as stated. ■

After showing that our sparsity functions are consistent, we prove that its maximizer $\widehat{\tau}_{x,h}$, in equation (17), is also consistent for τ_x.

Theorem 1. Under Assumptions A1 to A7, it holds that

$$\widehat{\tau}_{x,h} = \tau_x + o(h^{1/2}) + O_P\left(\frac{\log n}{nh}\right)^{1/4} \tag{22}$$

uniformly for $x \in \mathcal{X}$ and $h \in [\underline{h}_n, \bar{h}_n]$.

Proof. The proof of Theorem (1) consists of two parts: initially, it is proved that $\widehat{\tau}_{x,h}$ is consistent; then, in the second part of the proof, we calculate its rate of convergence.

Part 1 (T1): First, by the definition of $\widehat{\tau}_{x,h}$ and through Lemma 2, we have

$$\widehat{s}_{x,h}(\widehat{\tau}_{x,h}) \geqslant \widehat{s}_{x,h}(\tau_x) = s_x(\tau_x) + r_n,$$

where $r_n = o(h) + O_P\left(\sqrt{\log n/(nh)}\right)$ uniformly over τ, x and h. Hence,

$$\begin{aligned} s_x(\tau_x) - s_x(\widehat{\tau}_{x,h}) &\leqslant \widehat{s}_{x,h}(\widehat{\tau}_{x,h}) - s_x(\widehat{\tau}_{x,h}) - r_n \\ &\leqslant |\widehat{s}_{x,h}(\widehat{\tau}_{x,h}) - s_x(\widehat{\tau}_{x,h})| + |r_n| \\ &\leqslant \sup_{\tau,x,h} |\widehat{s}_{x,h}(\tau) - s_x(\tau)| + \sup_{\tau,x,h} |r_n| \\ &= o(h) + O_P\left(\sqrt{\log n/(nh)}\right) \end{aligned} \tag{23}$$

where the last equality follows again by Lemma 2.

Now, notice that compactness of $\mathcal{X}$, together with Assumptions A2, A6 and A7, ensure there exists an $\mathrm{x} \in \mathcal{X}$ such that

$$\sup_{x\in\mathcal{X}} \sup_{\tau:\ |\tau-\tau_x|\geqslant\epsilon} s_x(\tau) - s_x(\tau_x) = \sup_{\tau:\ |\tau-\tau_{\mathrm{x}}|\geqslant\epsilon} s_{\mathrm{x}}(\tau) - s_{\mathrm{x}}(\tau_{\mathrm{x}}) < 0$$

In view of this and using A6 once more, the following holds: for each $\epsilon > 0$ there exists an $\eta > 0$ such that the bound

$$s_x(\tau) \leqslant s_x(\tau_x) - \eta$$

holds for all x in the support of X and all τ with $|\tau - \tau_x| \geqslant \epsilon$.

Using compactness of $\mathcal{X} \times [\underline{h}_n, \bar{h}_n]$ and letting (x, h) attain the supremum $\sup_{x,h} |\widehat{\tau}_{x,h} - \tau_x|$ over $\mathcal{X} \times [\underline{h}_n, \bar{h}_n]$, we have

$$\{|\widehat{\tau}_{\mathrm{x,h}} - \tau_{\mathrm{x}}| \geqslant \epsilon\} \subseteq \{s_{\mathrm{x}}(\tau_{\mathrm{x}}) - s_{\mathrm{x}}(\widehat{\tau}_{\mathrm{x,h}}) \geqslant \eta\} \subseteq \{\sup_{x,h} s_x(\tau_x) - s_x(\widehat{\tau}_{x,h}) \geqslant \eta\}.$$

Thus, for any $\epsilon > 0$,

$$\mathbb{P}\{\sup_{x,h} |\widehat{\tau}_{x,h} - \tau_x| \geqslant \epsilon\} \leqslant \mathbb{P}\{\sup_{x,h} s_x(\tau_x) - s_x(\widehat{\tau}_{x,h}) \geqslant \eta\} \to 0$$

in view of (23).

Part 2 (T1) Recall the equations (13) and (15) with $D(\tau)$ and $\widehat{D}_h(\tau)$ defined as in (14) and (16). The first derivative of $s_x(\tau)$ is as:

$$s_x^{(1)}(\tau) := \frac{\partial s_x(\tau)}{\partial \tau} = x^\intercal \left[[D(\tau)]^{-1} D^{(1)}(\tau) [D(\tau)]^{-1} \mathbb{E} X \right]$$

with $D^{(1)}(\tau)$ defined as $\partial D(\tau)/\partial\tau = \mathbb{E}[XX^\intercal f^{(1)}(X^\intercal\beta(\tau)|X) \cdot X^\intercal [D(\tau)]^{-1}\mathbb{E}X]$; the derivation of $D^{(1)}(\tau)$ is found in Appendix A.2, equation (34).

The first order condition $s_x^{(1)}(\tau_x) = 0$ and its sample analog, $\widehat{s}^{(1)}(\widehat{\tau}_{x,h}) = 0$ yield:

$$\begin{aligned} x^\intercal [D(\tau_x)]^{-1} D^{(1)}(\tau_x) [D(\tau_x)]^{-1} \mathbb{E} X &= 0 \\ x^\intercal [\widehat{D}_h(\widehat{\tau}_{x,h})]^{-1} \widehat{D}^{(1)}(\widehat{\tau}_{x,h}) [D(\widehat{\tau}_{x,h})]^{-1} \bar{X} &= 0 \end{aligned}$$

Now, by a Taylor expansion with Lagrange remainder, we have

$$s_x(\widehat{\tau}_{x,h}) = s_x(\tau_x) + s_x^{(1)}(\tau_x)[\widehat{\tau}_{x,h} - \tau_x] + \frac{1}{2}s_x^{(2)}(\tau_x^*)[\widehat{\tau}_{x,h} - \tau_x]^2$$

with τ_x^* as a point between $\widehat{\tau}_{x,h}$ and τ_x. Assumptions A2, A6, and A7 ensure that $\tau \mapsto s_x(\tau)$ is strictly convex in a vicinity of τ_x, so $\inf_\tau s^{(2)}(\tau) > 0$ in such a vicinity.

Applying the first-order condition $s_x(\tau_x) = 0$, we can rewrite the expansion as:

$$s_x(\widehat{\tau}_{x,h}) = s_x(\tau_x) + \frac{1}{2}s_x^{(2)}(\tau_x^*)[\widehat{\tau}_{x,h} - \tau_x]^2, \tag{24}$$

which leads to

$$|\widehat{\tau}_{x,h} - \tau_x| = \sqrt{2}\sqrt{\frac{|s_x(\widehat{\tau}_{x,h}) - s_x(\tau_x)|}{|s_x^{(2)}(\tau_x^*)|}} = \sqrt{\frac{o(h) + O_P(\sqrt{\log n/(nh)})}{O_P(1)}}$$

Using $\sqrt{a+b} \leqslant \sqrt{a} + \sqrt{b}$, it yields our rate of convergence for $\widehat{\tau}_{x,h}$:

$$|\widehat{\tau}_{x,h} - \tau_x| = o(h^{1/2}) + O_P\left(\frac{\log n}{nh}\right)^{1/4} \tag{25}$$

as stated. ■

Now that the consistency for the quantile modes is proved and the rates of convergence are defined, we proceed to state the consistency for the estimator of the mode, $\widehat{m}_h(x)$. Our second theorem is constructed using previous results from this paper and from Fernandes, Guerre, and Horta (2021).

Theorem 2. If Assumptions A1 to A7 hold, then

$$\widehat{m}_h(x) = m(x) + o(h^{1/2}) + O_P\Big(\frac{\log n}{nh}\Big)^{1/4} \tag{26}$$

uniformly for $x \in \mathcal{X}$ and $h \in [\underline{h}_n, \bar{h}_n]$.

Proof. From Theorem 1 we know that $\widehat{\tau}_{x,h} \xrightarrow{p} \tau_x$ at a rate of $o(h^{1/2}) + O_P(\log n/nh)^{1/4}$. Also, from Theorems 1 and 2 in Fernandes, Guerre, and Horta (2021) we have:

$$\big\|\widehat{\beta}_h(\tau) - \beta(\tau)\big\| = O_P\left(\frac{1}{\sqrt{n}} + h^2\right) \tag{27}$$

Recalling that

$$m(x) = x^\intercal \beta(\tau_x) \quad \text{and} \quad \widehat{m}_h(x) = x^\intercal \widehat{\beta}_h(\widehat{\tau}_{x,h})$$

we obtain

$$\begin{aligned} |\widehat{m}_h(x) - m(x)| &= \left| x^\intercal \widehat{\beta}_h(\widehat{\tau}_{x,h}) - x^\intercal \beta(\tau_x) \right| \\ &= \left| x^\intercal \left(\widehat{\beta}_h(\widehat{\tau}_{x,h}) - \beta(\tau_x) \right) \right| \leqslant \|x\| \cdot \left\| \widehat{\beta}_h(\widehat{\tau}_{x,h}) - \beta(\tau_x) \right\|. \end{aligned} \tag{28}$$

Given the differentiability condition (A2), we have that $\beta(\tau)$ is Lipschitz-continuous, thus, for some constant $C > 0$, we have

$$\begin{aligned} \|\widehat{\beta}_h(\widehat{\tau}_{x,h}) - \beta(\tau_x)\| &\leqslant \|\widehat{\beta}_h(\widehat{\tau}_{x,h}) - \beta(\widehat{\tau}_{x,h})\| + \|\beta(\widehat{\tau}_{x,h}) - \beta(\tau_x)\| \\ &\leqslant (\sup_\tau \|\widehat{\beta}_h(\tau) - \beta(\tau)\|) + C\|\widehat{\tau}_{x,h} - \tau_x\| \\ &\leqslant O_p\left(\frac{1}{\sqrt{n}} + h^2\right) + C\left[o(h^{1/2}) + O_P\left(\frac{\log n}{nh}\right)^{1/4}\right], \end{aligned} \tag{29}$$

which yields (26). ■

Remark 1. Denote our rate of convergence from Theorem 2 as R_{CMR} and the rate from the estimator proposed by Zhang, Kato, and Ruppert (2023) as R_{ZKR},

$$\begin{aligned} R_{CMR} &= O_P\left(n^{-1/4}h^{-1/4}(\log n)^{1/4}\right) + o(h^{1/2}) \\ R_{ZKR} &= O_P\left(n^{-1/2}h^{-3/2}(\log n)^{1/2} + h^2\right) \end{aligned} \tag{30}$$

Neither rate is dimension dependable, thus free from the "curse of dimensionality"; under certain conditions on h, our rate, R_{CMR}, is marginally slower than R_{ZKR}. Apart from the presence of the deterministic term, both rates are similar, and the difference lies on the selection of the bandwidth parameter, h. Given a certain bandwidth, R_{ZKR} can achieve, at best, a rate of $(n \log n)^{-2/7}$, similar to the rate of Kemp and Santos-Silva (2012) and faster than the rate in Ota, Kato, and Hara (2019). Despite the differences in convergence rates, we attain uniformity with respect to design points of the covariates and the bandwidth.

Remark 2. Rewrite R_{ZKR} in (30) as

$$O_P\left(\left[\frac{\log n}{nh^3}\right]^{1/2} + h^2\right). \tag{31}$$

The ratio

$$\frac{(\log n/(nh))^{1/4}}{(\log n/(nh^3))^{1/2}} = \frac{nh^5}{\log n} \tag{32}$$

diverges to infinity under Assumption (viii) in Zhang, Kato, and Ruppert (2023), so under this assumption our estimator cannot achieve O_P rates faster than theirs. Nevertheless, under our weaker Assumption A5, we can make (32) go to zero, for example by taking $h = (n/\log n)^{-1/5}b$ with $b \to 0$ and $b^3(n/\log n)^{-2/5} \to \infty$. However, this particular choice for the bandwidth is not contemplated due to Zhang, Kato, and Ruppert's (2023) assumptions, but is enabled by condition A5.

Remark 3. Importantly, our estimator $\widehat{m}_h(x)$ attains the rate in Theorem 2 uniformly both in x and h; on the other hand, the representation in Proposition 1 of Zhang, Kato, and Ruppert (2023) is not uniform for the bandwidth. Obtaining uniformity in h can be useful for 3 types of bandwidth choices: (i) data-driven bandwidth choices, as in Fernandes, Guerre, and Horta (2021); (ii) adaptive bandwidth choices, such as the ones of Terrell and Scott (1992); Lepski et al. (1997); and (iii) choices robust to bandwidth-snooping, as in Armstrong and Kolesár (2018).

4 Concluding Remarks

In the present paper we developed a novel estimator for the conditional mode $\widehat{m}_h(x)$, called *Convolution Mode Regression*, based on inverting the smoothed quantile regression of Fernandes, Guerre, and Horta (2021). The idea of achieving the conditional mode via quantile regression is not groundbreaking, since it has been done previously (Ota, Kato, and Hara, 2019; Zhang, Kato, and Ruppert, 2023). Despite that, it presents advantages regarding the two main problems with mode regression, slow convergence in nonparametric settings and nonconvex optimization in linear environments. Our estimation strategy relies on an intermediate step in which, in order to estimate $\widehat{m}_h(x)$, we need to firstly estimate the conditional quantile function $Q(\cdot|x)$, and then the conditional quantile mode of Y given $X = x$, denoted τ_x. Thus, the mode estimation relies on estimating $Q(\cdot|x)$ and τ_x first.

Differently from the existing work of Zhang, Kato, and Ruppert (2023), who initially estimate the quantile regression then smooth it through a kernel, our approach relies on "smooth then estimate". We develop asymptotic consistency for our estimator, obtaining convergence rates similar to the ones of the estimator of Zhang, Kato, and Ruppert (2023), which, in the majority of cases, had a marginally faster rate. Apart from the initial smoothing, the main differentiation of our model from the authors' is in the bandwidth selection premise, since our assumption for the choice of h is less restrictive, without sacrificing significantly in terms of convergence rates. Furthermore, the uniformity $\widehat{m}_h(x)$ with respect to h makes our model an interesting choice when the bandwidth selection is data-driven or adaptive.

Further work related to present research can take many directions. In what we assess as more important, the continuation of the asymptotic properties, namely, the limiting distributions of the estimator. Furthermore, simulations similar to those Ongaratto and Horta (2021) did, comparing to the Ota, Kato, and Hara (2019) estimator, may be updated in order to evaluate the performance of the estimator against Zhang, Kato, and Ruppert's (2023). Comparisons with previous mode regression papers' econometric applications can be carried out, centered on assessing estimator performance. In accordance to that, a generalization of the present framework focused on time series can be done, with the interest in forecasting models for asymmetric data.

A Appendix

A.1 First and Second Derivatives of $\widehat{R}_h(b;\tau)$

From Fernandes, Guerre, and Horta (2021), the first and second derivatives of the smoothed sample objective function, $\widehat{R}_h(b;\tau)$, with respect to b, are, respectively:

$$\begin{aligned} \widehat{R}_h^{(1)}(b;\tau) &= \frac{1}{n}\sum_{i=1}^{n} X_i\left[K\left(\frac{X_i^\intercal b - Y_i}{h}\right) - \tau\right] \\ \widehat{R}_h^{(2)}(b;\tau) &= \frac{1}{n}\sum_{i=1}^{n} X_i X_i^\intercal k_h(X_i^\intercal b - Y_i) \end{aligned} \tag{33}$$

with $K(t) := \int_{-\infty}^{t} k(v)dv$.

A.2 Derivation of $D^{(1)}(\tau)$

Recalling the definition of $D(\tau)$:

$$D(\tau) := R^{(2)}(\beta(\tau);\tau) = \mathbb{E}[XX^\intercal f(X^\intercal \beta(\tau)|X)]$$

The first order differentiation is expressed as:

$$\begin{aligned} D^{(1)}(\tau) &:= \frac{\partial}{\partial\tau}\mathbb{E}[XX^\intercal f(X^\intercal\beta(\tau)|X)] \\ &= \mathbb{E}[XX^\intercal f^{(1)}(X^\intercal\beta(\tau)|X)\cdot X^\intercal \beta^{(1)}(\tau)] \\ &= \mathbb{E}[XX^\intercal f^{(1)}(X^\intercal\beta(\tau)|X)\cdot X^\intercal [D(\tau)]^{-1}\mathbb{E}X] \\ D^{(1)}(\tau) &= \mathbb{E}\Big[XX^\intercal f^{(1)}(X^\intercal\beta(\tau)|X)\cdot X^\intercal\Big\{\mathbb{E}[XX^\intercal f(X^\intercal\beta(\tau)|X)]\Big\}^{-1}\mathbb{E}X\Big] \end{aligned} \tag{34}$$

A.3 Derivation of $s_x(\tau)$

Recalling the definition of the population sparsity function:

$$s_x(\tau) = -\frac{\partial}{\partial\tau}x^\intercal\beta(\tau) = -x^\intercal \underbrace{[D(\tau)]^{-1}\mathbb{E}X}_{\beta^{(1)}(\tau)}$$

To calculate the first derivative of $s_x(\tau)$, we use the definition of $\beta(\tau)$ as in the previous equation:

$$s_x^{(1)}(\tau) := \frac{\partial s_x(\tau)}{\partial\tau} = \frac{\partial}{\partial\tau} - x^\intercal\beta^{(1)}(\tau) = -x^\intercal\beta^{(2)}(\tau) \tag{35}$$

Now, computing $\beta^{(2)}(\tau)$:

$$\begin{aligned}\beta^{(2)}(\tau) := \frac{\partial \beta^{(1)}(\tau)}{\partial \tau} &= \frac{\partial}{\partial \tau}[D(\tau)]^{-1}\mathbb{E}X \\ &= -D^{(1)}(\tau)[D(\tau)]^{-2}\mathbb{E}X \\ &= -D^{(1)}(\tau)[D(\tau)]^{-1}\beta^{(1)}(\tau) \\ &= -[D(\tau)]^{-1}D^{(1)}(\tau)[D(\tau)]^{-1}\mathbb{E}X\end{aligned} \tag{36}$$

Applying the result in (36) to equation (35) we get $s_x^{(1)}(\tau)$:

$$s_x^{(1)}(\tau) = x^\intercal\left[[D(\tau)]^{-1}D^{(1)}(\tau)[D(\tau)]^{-1}\mathbb{E}X\right] \tag{37}$$

with $D^{(1)}(\tau)$ defined as in equation (34).

The second derivative of $s_x(\tau)$ is required in the Taylor Expansion (24), so we compute $s^{(2)}(\tau)$ as follows:

$$\begin{aligned}s^{(2)}(\tau) := \frac{\partial s^{(1)}(\tau)}{\partial \tau} &= \frac{\partial}{\partial \tau}x^\intercal[D(\tau)]^{-1}D^{(1)}(\tau)[D(\tau)]^{-1}\mathbb{E}X \\ s^{(2)}(\tau) = x^\intercal &\left[\left([D(\tau)]^{-1}D^{(2)} - 2[D^{(1)}(\tau)]^2\right)[D(\tau)]^{-3}\right]\mathbb{E}X\end{aligned} \tag{38}$$

References

Armstrong, T. and M. Kolesár (2018). A simple adjustment for bandwidth snooping. *Review of Economic Studies 85*, 732–765.

Bassett, G. and R. Koenker (1982). An empirical quantile function for linear models with iid errors. *Journal of the American Statistical Association 77*(378), 407–415.

Chacón, J. E. (2020). The modal age of statistics. *International Statistical Review 88*(1), 122–141.

Chen, Y.-C. (2018). Modal regression using kernel density estimation: A review. *Wiley Interdisciplinary Reviewers: Computational Statistics 10*(4).

Chen, Y.-C., C. Genovese, R. Tibishirani, and L. Wasserman (2016). Nonparametric modal regression. *The Annals of Statistics 44*(2), 489–514.

Chernoff, H. (1964). Estimation of the mode. *Annals of the Institute of Statistical Mathematics 16*, 31–41.

Einbeck, J. and G. Tutz (2006). Modelling beyond regression functions: an application of multimodal regression to speed–flow data. *Applied Statistics 55*(4), 461–475.

Feng, Y., J. Fan, and J. Suykens (2020). A statistical learning approach to modal regression. *Journal of Machine Learning Research 21*(2), 1–35.

Fernandes, M., E. Guerre, and E. Horta (2021). Smoothing quantile regressions. *Journal of Business & Economic Statistics 39*(1), 338 – 357.

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

Kemp, G. C. R., P. M. D. C. Parente, and J. M. C. Santos-Silva (2020). Dynamic vector mode regression. *Journal of Business & Economic Statistics 38*(3), 647–661.

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

Koenker, R. and G. Bassett (1978). Regression quantiles. *Econometrica 46*(1), 33–50.

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

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

Lepski, O. V., E. Mammen, and V. G. Spokoiny (1997). Optimal spatial adaptation to inhomogeneous smoothness: an approach based on kernel estimates with variable bandwidth selectors. *The Annals of Statistics 25*(3), 929–947.

Liu, J., R. Zhanga, W. Zhaoa, and Y. Lv (2013). A robust and efficient estimation method for single index models. *Journal of Multivariate Analysis 122*, 226–238.

Matzner-Løfber, E., A. Gannoun, and J. G. D. Gooijer (1998). Nonparametric forecasting: a comparison of three kernel-based methods. *Communications in Statistics - Theory and Methods 27*(7), 1593–1617.

Nadaraya, E. A. (1964). Some new estimates for distribution functions. *Theory of Probability & Its Applications 9*(3), 497–500.

Newey, W. and D. McFadden (1994). Large sample estimation and hypothesis testing. In *Handbook of Econometrics (Vol. 4)*. Elselvier.

Ongaratto, A. and E. Horta (2021). Conditional mode: An approach via smoothed quantile regression. Master's thesis, Federal University of Rio Grande do Sul (UFRGS), https://lume.ufrgs.br/handle/10183/237727.

Ota, H., K. Kato, and S. Hara (2019). Quantile regression approach to conditional mode estimation. *Electronic Journal of Statistics 13*(2), 3120–3160.

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

Parzen, E. (1979). Nonparametric statistical data modeling. *Journal of the American Statistical Association 74*(365), 105–121.

Sager, T. (1978). Estimation of a multivariate mode. *The Annals of Statistics 6*(4), 802–812.

Sager, T. W. and R. A. Thisted (1982). Maximum likelihood estimation of isotonic modal regression. *The Annals of Statistics 10*(3), 690–707.

Scott, D. W. (1992). *Multivariate Density Estimation: Theory, practice and Visualization*. John Wiley & Sons.

Terrell, G. and D. Scott (1992). Variable kernel density estimation. *The Annals of Statistics 20*(3), 1236–1265.

Ullah, A., T. Wang, and W. Yao (2021). Modal regression for fixed effects panel data. *Empirical Economics 60*, 261–308.

van der Vaart, A. (1998). *Asymptotic Statistics*. Cambridge University Press.

Wang, T. (2024). Non-parametric estimator for conditional mode with parametric features. *Oxford Bulletin of Economics and Statistics 86*(1), 44–73.

Wang, X., H. Chen, W. Cai, D. Shen, and H. Huang (2017). Regularized modal regression with applications in cognitive impairment prediction. *Advances in Neural Information Processing Systems 30 (NIPS 2017)*.

Yao, W. and L. Li (2014). A new regression model: Modal linear regression. *Scandinavian Journal of Statistics 41*, 656–671.

Yao, W., B. G. Lindsay, and R. Li (2012). Local mode regression. *Journal of Nonparametric Statistics 24*(3), 647–663.

Yu, K. and K. Aristodemou (2012). Bayesian mode regression. *arXiv*.

Zhang, R., W. Zhao, and J. Liu (2013). Robust estimation and variable selection for semiparametric partially linear varying coefficient model based on modal regression. *Journal of Nonparametric Statistics 25*(2), 523–544.

Zhang, T., K. Kato, and D. Ruppert (2023). Bootstrap inference for quantile-based modal regression. *Journal of the American Statistical Association 118*(541), 122–134.