EMPIRICAL LIKELIHOOD FOR NONPARAMETRIC ADDITIVE MODELS

By

Taisuke Otsu

April 2011

COWLES FOUNDATION DISCUSSION PAPER NO. 1792

COWLES FOUNDATION FOR RESEARCH IN ECONOMICS
YALE UNIVERSITY
Box 208281
New Haven, Connecticut 06520-8281

http://cowles.econ.yale.edu/

Empirical Likelihood for Nonparametric Additive Models

Taisuke Otsu[*][†]
Cowles Foundation and Department of Economics
Yale University

April 2011

Abstract

Nonparametric additive modeling is a fundamental tool for statistical data analysis which allows flexible functional forms for conditional mean or quantile functions but avoids the curse of dimensionality for fully nonparametric methods induced by high-dimensional covariates. This paper proposes empirical likelihood-based inference methods for unknown functions in three types of nonparametric additive models: (i) additive mean regression with the identity link function, (ii) generalized additive mean regression with a known non-identity link function, and (iii) additive quantile regression. The proposed empirical likelihood ratio statistics for the unknown functions are asymptotically pivotal and converge to chi-square distributions, and their associated confidence intervals possess several attractive features compared to the conventional Wald-type confidence intervals.

1 Introduction

Nonparametric additive modeling is a fundamental tool for statistical data analysis which allows flexible functional forms for conditional mean or quantile functions but avoids the curse of dimensionality for fully nonparametric methods induced by high-dimensional covariates (see, e.g., Hastie and Tibshirani, 1990). This paper proposes empirical likelihood-based inference methods for unknown functions in nonparametric additive models.[1] In particular, we consider three types of the additive models: (i) additive mean regression with the identity link function, (ii) generalized additive mean regression with a known non-identity link function, and (iii) additive quantile regression. For these models, we find localized versions of estimating equations to estimate the unknown functions at given values of covariates, and construct empirical likelihood functions based on these estimating equations. The proposed empirical likelihood ratio statistics are asymptotically pivotal and converge to chi-square distributions. In other words, we can still observe the so-called Wilks phenomena (i.e., convergence of a likelihood ratio statistic

[*] E-mail: taisuke.otsu@yale.edu. Website: http://cowles.econ.yale.edu/faculty/otsu.htm. Address: P.O. Box 208281, New Haven, CT 06520-8281, USA. Phone: +1-203-432-9771. Fax: +1-203-432-6167.

[†] The author would like to thank anonymous referees for helpful comments.

[1] See Owen (2001) and Kitamura (2007) for a comprehensive review on empirical likelihood.

to the chi-square distribution) in these nonparametric additive models. Also, the confidence intervals obtained by inverting the empirical likelihood ratio statistics possess several attractive features compared to the conventional Wald-type confidence intervals, such as circumvention of asymptotic variance estimation to compute the standard error, flexible shapes of the confidence intervals determined by data, transformation invariance, and range-preserving property.

There is rich literature on statistical theory of nonparametric additive models. For the additive mean regression with the identity link function, Stone (1994) and Newey (1997) studied properties of series estimators. Buja, Hastie and Tibshirani (1989) and Hastie and Tibshirani (1990) developed the backfitting procedure which iteratively estimates each additive nonparametric component. Opsomer and Ruppert (1997) and Opsomer (2000) studied asymptotic properties of the backfitting estimator in depth. Mammen, Linton and Nielsen (1999) proposed a modification of backfitting, called smooth backfitting, to achieve better asymptotic properties such as oracle efficiency. On the other hand, Linton and Nielsen (1995) developed the marginal integration technique, which utilizes an integral expression for the unknown function of interest. Linton (1997) and Fan, Mammen and Härdle (1998) studied oracle efficiency of the marginal integration estimator. Horowitz, Klemelä and Mammen (2006) analyzed optimal properties of different estimators in a unified framework.

For the generalized additive mean regression with a known non-identity link function, Linton and Härdle (1996) extended the marginal integration approach of Linton and Nielsen (1995) to this context. Horowitz and Mammen (2004) developed a two stage estimation procedure, in which we first obtain a preliminary estimator for unknown functions based on series approximations and then refine the preliminary estimator by the second stage local polynomial fitting. Horowitz and Mammen's (2004) estimator is asymptotically normal and oracle efficient and achieves the optimal convergence rate derived by Stone (1985, 1986). Our construction of empirical likelihood is based on an estimating equation implied from the second stage local linear regression of Horowitz and Mammen (2004) by using the first stage preliminary estimator as inputs. Also Horowitz (2001) proposed a nonparametric estimator for the case where the link function is unknown.

For the additive quantile regression, Doksum and Koo (2000) studied a series estimation procedure and Goojier and Zerom (2003) extended the marginal integration approach to the quantile regression setup. Horowitz and Lee (2005) extended the two stage approach of Horowitz and Mammen (2004) to the quantile regression context and derived analogous optimal properties to the mean regression case. Our construction of empirical likelihood utilizes an estimating equation implied from the second stage local quantile regression of Horowitz and Lee (2005).

This paper also contributes to the rapidly growing literature on empirical likelihood (Owen, 1988, 2001). Compared to inference problems for parametric or finite-dimensional components (e.g., Wang and Jing, 2003; Otsu, 2007; Hjort, McKeague and van Keilegom, 2009), the literature on empirical likelihood inference for nonparametric or infinite-dimensional components is relatively thin. Chen and Qin (2000) proposed an empirical likelihood confidence interval for the conditional mean function based on an estimating equation of local linear fitting, and showed that their empirical likelihood confidence interval has better higher-order coverage properties than the Wald-type confidence interval. This paper can be considered as an extension of Chen and Qin's (2000) approach to nonparametric additive models. Fan, Zhang and Zhang (2001) provided several nonparametric settings where we can observe the Wilks phenomena. This paper provides additional positive results for the Wilks phenomena in nonparametric additive models.

This paper is organized as follows. Section 2 considers the nonparametric additive mean regression model with the identity link function, proposes the empirical likelihood function for an unknown function, and studies its asymptotic property. Section 3 discusses the case of the generalized additive mean regression model with a known non-identity link function. Section 4 extends the empirical likelihood approach to the nonparametric additive quantile regression model. Section 5 concludes. All proofs and lemmas are contained in the Appendix.

2 Additive Mean Regression with Identity Link Function

The notation closely follows that of Horowitz and Mammen (2004). We first consider the nonparametric additive regression model with the identify (or linear) link function:

$$Y = \mu + m_1(X^1) + \cdots + m_d(X^d) + U, \tag{1}$$

$$E\left[U \mid X = x\right] = 0 \text{ for a.e. } x,$$

where $Y \in \mathbb{R}$ is a scalar response variable, $X^j \in \mathbb{X}_j \subset \mathbb{R}$ $(j = 1, \ldots, d)$ is a scalar explanatory random variable, $X = \left(X^1, \ldots, X^d\right)'$, $U \in \mathbb{R}$ is an unobservable error term satisfying the mean independence condition $E\left[U \mid X = x\right] = 0$ for almost every x, μ is an unknown constant, and $m_j : \mathbb{X}_j \to \mathbb{R}$ $(j = 1, \ldots, d)$ is an unknown function. Note that this model is more restrictive than the fully nonparametric regression (i.e., $Y = m\left(X^1, \ldots, X^d\right) + \epsilon$) due to the additive structure. However, the additive regression (1) provides an attractive compromise between fully parametric and nonparametric models since the convergence rates of nonparametric estimators for m_j's typically do not increase with the number of covariates d (i.e., avoid the curse of dimensionality).

To simplify the presentation and technical discussion, hereafter we assume that the support of X^j is $\mathbb{X}_j = [-1, 1]$ for all $j = 1, \ldots, d$, and normalize m_j's as $\int_{-1}^{1} m_j(v)dv = 0$ for all $j = 1, \ldots, d$. Based on an i.i.d sample $\{Y_i, X_i\}_{i=1}^n$, we wish to conduct inference on the unknown function $m_1(x^1)$ evaluated at some $x^1 \in \mathbb{X}_1$. Inference on the other components $m_j(x^j)$ $(j = 2, \ldots, d)$ can be implemented in the same manner.

The nonparametric additive regression model (1) and its generalizations discussed in the following sections are typically applied when the dimension of the explanatory variables X is large. In this case, since it is difficult to visualize the estimates of the whole regression function $\mu + m_1(x^1) + \cdots + m_d(x^d)$, we commonly report the plots for the estimates of $m_j(x^j)$ $(j = 1, \ldots, d)$ separately. Therefore, the confidence interval for $m_j(x^j)$, which is plotted along the estimates of $m_j(x^j)$, is a fundamental tool to evaluate the uncertainty of the estimates of $m_j(x^j)$ and to assess the functional form of the regression function. For empirical applications of nonparametric additive regression, see, e.g., Fan and Jiang (2005) (additive mean regression for housing price in Boston) and Horowitz and Lee (2005) (additive median regression for sales of Japanese firms in the chemical industry). Also Hastie and Tibshirani (1990) contain various real data examples of nonparametric additive regression. In these examples, most estimation results are presented by separate plots for the estimates of $m_j(x^j)$'s, where our confidence intervals discussed below can be added along the estimates.

To construct the empirical likelihood function for the object of interest $m_1(x^1)$, let us tentatively assume that the other functions $m_2, \ldots, m_d$ and the intercept μ are known. Then the variable $Y^* = Y - \mu - m_2(X^2) - \cdots - m_d(X^d)$ is observable and we can identify the object of interest $m_1(x^1)$ by the conditional mean $m_1(x^1) = E\left[Y^* \middle| X^1 = x^1\right]$. Thus, we can estimate $m_1(x^1)$ by, for example, the local linear regression, where we solve the weighted least square problem

$$\min_{a,b} \sum_{i=1}^{n} K_h\left(x^1 - X_i^1\right)\left\{Y_i^* - a - b\left(X_i^1 - x^1\right)\right\}^2, \tag{2}$$

and estimate $m_1(x^1)$ by the solution $\hat{a}$ with respect to a. Here $K_h(v) = K(v/h)$ with a kernel function K and a bandwidth parameter h satisfying $h \to 0$ as $n \to \infty$. After some manipulation, we can see that the solution $\hat{a}$ satisfies the first-order condition (see, Fan and Gijbels, 1996)

$$\sum_{i=1}^{n} \tilde{K}_i\left(Y_i^* - \hat{a}\right) = 0,$$

where

$$\tilde{K}_i = K_h\left(x^1 - X_i^1\right)\left\{\begin{array}{c} \frac{1}{nh}\sum_{j=1}^n K_h\left(x^1 - X_j^1\right)\left(\frac{X_j^1 - x^1}{h}\right)^2 \\ -\left(\frac{X_i^1 - x^1}{h}\right)\frac{1}{nh}\sum_{j=1}^n K_h\left(x^1 - X_j^1\right)\left(\frac{X_j^1 - x^1}{h}\right) \end{array}\right\}.$$

If we regard this condition as an estimating equation for the expectation $E[\hat{a}]$, the empirical likelihood function for $E[\hat{a}]$ can be defined as

$$L^*(a) = \sup_{\{p_i\}_{i=1}^n} \prod_{i=1}^{n} p_i, \tag{3}$$

$$\text{s.t. } 0 \leq p_i \leq 1,\ \sum_{i=1}^{n} p_i = 1,\ \sum_{i=1}^{n} p_i \tilde{K}_i\left(Y_i^* - a\right) = 0.$$

Note that without the last constraint $\sum_{i=1}^n p_i \tilde{K}_i\left(Y_i^* - a\right) = 0$, the above supremum becomes n^{-n}. Thus, the (normalized) log empirical likelihood ratio is defined as $\ell^*(a) = -2\left\{\log L^*(a) + n\log n\right\}$. Although the optimization problem (3) involves n variables $\{p_i\}_{i=1}^n$, mild regularity conditions allow an application of the Lagrange multiplier method (see, e.g., Theorem 2.2 in Newey and Smith, 2004), and the dual form for $\ell^*(a)$ is written as

$$\ell^*(a) = 2 \sup_{\lambda \in \Lambda_n^*(a)} \sum_{i=1}^{n} \log\left(1 + \lambda \tilde{K}_i\left(Y_i^* - a\right)\right), \tag{4}$$

where $\Lambda_n^*(a) = \left\{\lambda \in \mathbb{R} : \lambda \tilde{K}_i(Y_i^* - a) \in \mathbb{V}^* \text{ for } i = 1, \ldots, n\right\}$ and $\mathbb{V}^*$ is an open interval containing 0.

In practice, we use the dual representation (4) instead of the original problem (3) to compute the empirical likelihood function. Note that the optimization problem for the Lagrange multiplier λ in (4) is one-dimensional, and that the objective function $\sum_{i=1}^n \log\left(1 + \lambda \tilde{K}_i(Y_i^* - a)\right)$ for λ is typically concave in λ. Therefore, the conventional Newton-type gradient-based optimization routine can be used to evaluate the empirical likelihood ratio $\ell^*(a)$.

Note that the above construction of $\ell^*(a)$ gives us the empirical likelihood ratio for the expectation $E[\hat{a}]$, rather than for the object of interest $m_1(x^1)$ itself. However, if we choose a relatively fast decay rate for the bandwidth h (i.e., undersmoothing), then the bias component $m_1(x^1) - E[\hat{a}]$ becomes asymptotically negligible. Therefore, the function (4) can be employed as a valid empirical likelihood ratio for $m_1(x^1)$.

If we observe Y^*, we can see that the empirical likelihood ratio $\ell^*(a)$ evaluated at $a = m_1(x^1)$ converges in distribution to the χ_1^2 distribution under suitable regularity conditions (see, Chen and Qin, 2000). Thus, the asymptotic confidence interval for $m_1(x^1)$ can be constructed by inverting the empirical likelihood ratio test statistic, i.e., $\left\{a : \ell^*(a) \leq \chi^2_{1,1-\alpha}\right\}$, where $\chi^2_{1,1-\alpha}$ is the $100(1-\alpha)\%$ critical value for the χ_1^2 distribution. However, in practice, we do not observe Y^* since m_j's and μ are unknown. Therefore, we find a proxy for Y^* by utilizing the first stage preliminary estimation of $m_2, \ldots, m_d$ and μ in Horowitz and Mammen (2004) and propose a feasible version of the empirical likelihood function $\ell^*(a)$.

To obtain the first stage estimator for m_j's having a sufficiently fast convergence rate, Horowitz and Mammen (2004) employed a series estimator. Consider a basis $\{p_k\}_{k=1}^{\infty}$ for smooth functions on $[-1, 1]$, which satisfies

$$m_j(x^j) = \sum_{k=1}^{\infty} \theta_{jk} p_k(x^j),$$

for all $x^j \in [-1, 1]$ and $j = 1, \ldots, d$, and some coefficients $\{\theta_{jk}\}$. Also assume that the basis is orthogonal (i.e., $\int_{-1}^{1} p_j(v) p_k(v)\, dv = I\{j = k\}$) and satisfies a normalization constraint $\int_{-1}^{1} p_k(v)\, dv = 0$. If we truncate the infinite series representation for m_j's by a positive integer κ (satisfying $\kappa \to \infty$ as $n \to \infty$), a series approximation for $\mu + m_1(x^1) + \cdots + m_d(x^d)$ is obtained as $P_\kappa(x)' \theta_\kappa$ for some $\theta_\kappa \in \mathbb{R}^{\kappa d + 1}$, where $P_\kappa(x) = \left[1, p_1(x^1), \ldots, p_\kappa(x^1), \ldots, p_1(x^d), \ldots, p_\kappa(x^d)\right]'$. If we estimate the coefficients θ_κ by the least square method

$$\hat{\theta}_\kappa = \arg\min_{\theta_\kappa} \sum_{i=1}^{n} \left\{Y_i - P_\kappa(X_i)' \theta_\kappa\right\}^2,$$

then the unknown function $\mu + m_1(x^1) + \cdots + m_d(x^d)$ can be estimated by $P_\kappa(x)' \hat{\theta}_\kappa$. Note that since this series estimator $P_\kappa(x)' \hat{\theta}_\kappa$ imposes the additive structure in the original model (1), it does not involve any higher dimensional nonparametric estimation, which enables us to avoid the curse of dimensionality.

Horowitz and Mammen (2004) used the series estimator $P_\kappa(x)' \hat{\theta}_\kappa$ as inputs to the second stage point estimation of $m_1(x^1)$. We employ this estimator to construct a feasible empirical likelihood function for inference on $m_1(x^1)$. Note that the intercept μ is estimated by the first component of $\hat{\theta}_\kappa$ (denote by $\tilde{\mu}$) and the function $m_j(x^j)$ is estimated by an adequate component of $P_\kappa(x)' \hat{\theta}_\kappa$ (denote by $\tilde{m}_j(x^j)$). Then a feasible analog of $Y_i^* = Y_i - \mu - m_2(X_i^2) - \cdots - m_d(X_i^d)$ is defined as

$$\tilde{Y}_i = Y_i - \tilde{\mu} - \tilde{m}_2(X_i^2) - \cdots - \tilde{m}_d(X_i^d). \tag{5}$$

By replacing Y_i^* in (4) with its proxy $\tilde{Y}_i$, we propose the following feasible empirical likelihood function:

$$\ell(a) = 2 \sup_{\lambda \in \Lambda_n(a)} \sum_{i=1}^n \log\left(1 + \lambda \tilde{K}_i\left(\tilde{Y}_i - a\right)\right), \tag{6}$$

where $\Lambda_n(a) = \left\{\lambda \in \mathbb{R} : \lambda \tilde{K}_i\left(\tilde{Y}_i - a\right) \in \mathbb{V} \text{ for } i = 1, \ldots, n\right\}$ and $\mathbb{V}$ is an open interval containing 0.

Under similar assumptions of Horowitz and Mammen (2004) listed in Appendix A.1, the asymptotic property of the empirical likelihood ratio $\ell(a)$ evaluated at $a = m_1(x^1)$ is obtained as follows.

Theorem 2.1. *Under Assumptions 1-4 in Appendix A.1,*

$$\ell\left(m_1(x^1)\right) \xrightarrow{d} \chi_1^2,$$

for each $x^1 \in [-1, 1]$.

Remark 2.1 (Intuition for technical argument)**.** The assumptions for this theorem are adaptations of Horowitz and Mammen (2004, Assumptions A1-A7) to the present setting, where the link function is identity. In contrast to Horowitz and Mammen (2004), we impose undersmoothing $nh^5 \to 0$ for the bandwidth h (Assumption 4(ii)) to neglect an asymptotic bias component. If we set $h = Cn^{-1/5}$ as in Horowitz and Mammen (2004), the empirical likelihood ratio $\ell\left(m_1(x^1)\right)$ converges to a non-central χ_1^2 distribution. Intuitively, under our assumptions, the series estimator $\tilde{m}_j$ converges to m_j at a sufficiently fast rate and thus the proxy $\tilde{Y}_i$ is sufficiently close to Y_i^*. Therefore, we can establish the asymptotic equivalence between the empirical likelihood ratio $\ell\left(m_1(x^1)\right)$ and its infeasible version $\ell^*\left(m_1(x^1)\right)$, and a modified argument of Chen and Qin (2000) and Otsu and Xu (2010) implies that $\ell^*\left(m_1(x^1)\right)$ has the χ_1^2 limiting distribution. Also, in the context of point estimation for $m_1(x^1)$, Horowitz and Mammen (2004) showed a so-called oracle property: the local linear regression from $\tilde{Y}_i$ on X_i^1 has the same first-order asymptotic property as the one from Y_i^* on X_i^1. Theorem 2.1 can be considered as an analog of the oracle property to the empirical likelihood context.

Remark 2.2 (Wilks phenomenon)**.** Theorem 2.1 says that the empirical likelihood ratio $\ell\left(m_1(x^1)\right)$ is asymptotically pivotal and converges to the χ_1^2 distribution, i.e., the Wilks phenomenon emerges in the context of nonparametric additive regression. This result can be compared with earlier works which also have demonstrated the Wilks phenomenon for empirical likelihood in other nonparametric models, such as Chen and Qin (2000), Fan, Zhang and Zhang (2001), and Otsu and Xu (2010). Intuitively, the moment restriction $E\left[\tilde{K}_i\left(\tilde{Y}_i - m_1(x^1)\right)\right] \approx 0$ can be viewed as a "localized" moment restriction at $X_i^1 = x^1$ with an effective sample size nh, instead of n for standard moment restrictions. By undersmoothing the tuning parameters h and κ, we can neglect the bias in $E\left[\tilde{K}_i\left(\tilde{Y}_i - m_1(x^1)\right)\right]$ from 0, and an adaptation of a standard argument from the empirical likelihood literature for estimating equations (e.g., Qin and Lawless, 1994) implies the Wilks phenomenon in our nonparametric context.

Remark 2.3 (Confidence interval)**.** Based on Theorem 2.1, the $100\,(1-\alpha)\,\%$ asymptotic empirical likelihood confidence interval for $m_1(x^1)$ is obtained by inverting the empirical likelihood ratio statistic $\ell\left(m_1(x^1)\right)$, i.e.,

$$ELCI_\alpha = \left\{a : \ell\,(a) \leq \chi^2_{1,1-\alpha}\right\}.$$

Compared to the Wald-type confidence interval (i.e., the point estimate±2×standard error), there are at least four advantages for the empirical likelihood confidence interval. First, the empirical likelihood confidence interval does not require the estimation of the asymptotic variance, which typically involves additional nonparametric estimation for the conditional variance $Var\left(U|\, X^1 = x^1\right)$ and the marginal density $f_1\left(x^1\right)$ of X^1. In the next remark, we argue that in some special case this circumvention of variance estimation can yield a better higher-order coverage property for the empirical likelihood confidence interval. Second, the empirical likelihood confidence interval is not necessarily symmetric around the point estimator of $m_1(x^1)$, i.e., the shape of the confidence interval is determined by that of the empirical likelihood function. Intuitively, the Wald-type confidence interval is derived from a quadratic approximation to some criterion function to estimate $m_1(x^1)$. The empirical likelihood confidence interval is derived directly from the empirical likelihood function without relying on such a quadratic approximation. Third, the empirical likelihood confidence interval is transformation invariant, i.e., based on $ELCI_\alpha$, the $100\,(1-\alpha)\,\%$ asymptotic confidence interval for a transformed object $q\left(m_1(x^1)\right) \in \mathbb{R}$ is obtained as $\{q\,(a) : a \in ELCI_\alpha\}$. Finally, the empirical likelihood confidence interval is range-preserving, i.e., if the value of $m_1(x^1)$ is restricted to a subset $\mathbb{M}$ of $\mathbb{R}$ (e.g., $m_1(x^1) \geq 0$), then $ELCI_\alpha$ is always a subset of $\mathbb{M}$ because we set $\ell\,(a) = \infty$ for any $a \in \mathbb{R} \setminus \mathbb{M}$.

Remark 2.4 (Higher-order property)**.** We present some intuition for why the empirical likelihood confidence interval can be theoretically better than the Wald-type confidence interval. Assume that the functions $m_2, \ldots, m_d$ and the intercept μ are known and consider the (infeasible) empirical likelihood function $\ell^*\,(a)$ defined in (4). The same argument to Theorem 2.1 yields $\ell^*\left(m_1(x^1)\right) \xrightarrow{d} \chi^2_1$, and the associated empirical likelihood confidence interval for $m_1(x^1)$ is defined as $ELCI^*_\alpha = \left\{a : \ell^*\,(a) \leq \chi^2_{1,1-\alpha}\right\}$. On the other hand, the Wald-type confidence interval for $m_1(x^1)$ based on the local linear estimator $\hat{a}$ obtained from the solution of (2) is defined as $WCI^*_\alpha = \left[\hat{a} \pm z_{1-\alpha/2}\sqrt{\widehat{Asy.Var}\,(\hat{a})}\right]$, where $\widehat{Asy.Var}\,(\hat{a})$ is a nonparametric estimator for the asymptotic variance of $\hat{a}$ and $z_{1-\alpha/2}$ is the $(1-\alpha/2)$-th quantile of the standard normal distribution. Under this setup with additional regularity conditions, we can directly apply the results of Chen and Qin (2000). Chen and Qin (2000) found that even though both $ELCI^*_\alpha$ and WCI^*_α are derived from the local linear regression problem in (2), their coverage errors for

$m_1(x^1)$ have different orders near the boundary of the support $[-1,1]$ for X^1, i.e.,

$$\Pr\left\{m_1(x^1) \in ELCI^*_\alpha\right\} = 1 - \alpha + O\left(nh^5 + h^2 + (nh)^{-1}\right),$$
$$\Pr\left\{m_1(x^1) \in WCI^*_\alpha\right\} = 1 - \alpha + O\left(nh^5 + h + (nh)^{-1}\right),$$

for all $x^1 \in [-1,-1+h] \cup [1-h,1]$.[2] For example, if $h = O\left(n^{-1/3}\right)$, then the coverage error of $ELCI^*_\alpha$ is $O\left(n^{-2/3}\right)$ but the coverage error of WCI^*_α is $O\left(n^{-1/3}\right)$. As Chen and Qin (2000) argued, this higher-order difference near the boundary emerges from the fact that the coverage error of WCI^*_α depends on the estimation error of the asymptotic variance of $\hat{a}$. Since the empirical likelihood confidence interval is free from such an estimation error, $ELCI^*_\alpha$ yields a better higher-order coverage property than WCI^*_α near the boundary of the support.[3] The analysis for the (feasible) empirical likelihood ratio $\ell(a)$ in (6) is considerably more complicated because of the first stage estimation of μ and m_j's. Therefore, formal higher-order analysis is beyond the scope of the paper. However, it is reasonable to expect that similar arguments to Chen and Qin (2000) will yield analogous higher-order properties.

Remark 2.5 (Practical consideration)**.** To compute the empirical likelihood ratio statistic $\ell(a)$, we need to choose the basis $\{p_k\}$, series length κ, kernel function K, and bandwidth h. The assumptions on the basis (Assumption 3 in Appendix A.1) are standard and satisfied by popular basis functions, such as Fourier and spline bases. To choose the series length κ for the first stage estimation of μ and m_j's, we can apply conventional methods, such as cross validation, to control the estimation error (see, e.g., Chen, 2007). The assumptions on the kernel function K (Assumption 4 in Appendix A.1) are also mild and allow popular density functions, such as the uniform, triangular, and Epanechnikov. For the bandwidth parameter h, note that Assumption 4 (ii) in Appendix A.1 requires undersmoothing (i.e., $nh^5 \to 0$) which prohibits direct applications of the plug-in and penalized least square methods proposed by Horowitz and Mammen (2004). Also, it is not clear whether the bandwidth selection procedures by Horowitz and Mammen (2004), which intend to minimize the mean squared error for point estimation of $m_1(x^1)$, yield desirable coverage properties for the confidence interval of $m_1(x^1)$. Instead we suggest to employ a plug-in approach based on the optimal bandwidth derived by Chen and Qin (2000), which minimizes the leading coverage error of the empirical likelihood confidence interval for the conditional mean. In particular, we consider an auxiliary nonparametric regression from $\tilde{Y}_i$ on X^1_i and estimate Chen and Qin's (2000) optimal bandwidth ("h^*" in their notation) by taking the sample analogs. Since Chen and Qin's (2000) optimal bandwidth is of order $O\left(n^{-1/3}\right)$, this choice satisfies the undersmoothing condition, $nh^5 \to 0$.

Remark 2.6 (Inference on derivatives)**.** Although this paper focuses on inference for the regression function $m_1(x^1)$, it is possible to extend our empirical likelihood approach to conduct inference on the derivative $m_1'(x^1) = dm_1(x^1)/dx^1$. In the additive model (1), the derivative $m_1'(x^1)$ gives us the

[2] In the interior of the support, both $ELCI^*_\alpha$ and WCI^*_α have coverage errors of the same order $O\left(nh^5 + h^2 + (nh)^{-1}\right)$.

[3] Chen and Qin (2000) also proposed Bartlett correction for $ELCI^*_\alpha$, which provides even smaller coverage errors.

marginal effect $\partial E\left[Y|\, X=x\right]/\partial x^1$ for x^1. Observe that $m_1'(x^1)$ is estimated by the solution of the local linear regression in (2) with respect to b, and the solution $\hat{b}$ satisfies the first-order condition

$$\sum_{i=1}^{n} \bar{K}_i \left(Y_i^* - \hat{b}\left(X_i^1 - x^1\right)\right) = 0,$$

where

$$\bar{K}_i = K_h\left(x^1 - X_i^1\right)\left\{\frac{1}{nh}\sum_{j=1}^{n} K_h\left(x^1 - X_j^1\right)\left(\frac{X_j^1 - x^1}{h}\right) - \left(\frac{X_i^1 - x^1}{h}\right)\frac{1}{nh}\sum_{j=1}^{n} K_h\left(x^1 - X_j^1\right)\right\}.$$

Therefore, by using $\tilde{Y}_i$ defined in (5), the (dual) empirical likelihood function for $m_1'(x^1)$ can be defined as

$$\ell_1\left(b\right) = 2\sup_{\lambda\in\Lambda_{1n}(b)}\sum_{i=1}^{n}\log\left(1+\lambda\bar{K}_i\left(\tilde{Y}_i - b\left(X_i^1 - x^1\right)\right)\right),$$

$\Lambda_{1n}\left(b\right) = \left\{\lambda\in\mathbb{R} : \lambda\bar{K}_i\left(\tilde{Y}_i - b\left(X_i^1 - x^1\right)\right)\in\mathbb{V}_1 \text{ for } i=1,\ldots,n\right\}$ and $\mathbb{V}_1$ is an open interval containing 0. Based on Qin and Tsao (2005), we conjecture that the empirical likelihood ratio $\ell_1\left(m_1'(x^1)\right)$ will converge in distribution to a scaled χ^2 distribution. It is interesting to extend this approach to higher-order derivatives by considering estimating equations for higher-order local polynomial regressions.

Remark 2.7 (Significance test)**.** Also Theorem 2.1 can be employed as a basis for hypothesis testing on the additive regression model. For example, if we want to test $H_0 : m_1(x^1) = 0$ against $H_1 : m_1(x^1) \neq 0$ at some given x^1, we can use the empirical likelihood ratio statistic $\ell\left(0\right)$. To test the overall significance of X^1 over a subset $\mathbb{S}\subset[-1,1]$, the researcher may be interested in testing $H_0 : m_1(x^1) = 0$ for all $x^1\in\mathbb{S}$ against $H_1 : m_1(x^1)\neq 0$ for some $x^1\in\mathbb{S}$. In this case, we can consider the integrated test statistic $\int_{x^1\in\mathbb{S}}\ell\left(0;x^1\right)dx^1$, where $\ell\left(0;x^1\right)$ is the empirical likelihood ratio statistic $\ell\left(0\right)$ evaluated at $x^1\in\mathbb{S}$. This approach is adopted by Chen, Härdle and Li (2003) to test goodness-of-fit for a parametric model. Although it is not a focus of this paper, it is interesting to investigate statistical properties of this test statistic.

3 Additive Mean Regression with Non-Identity Link Function

We next consider the nonparametric generalized additive regression model with a non-identify (or non-linear) link function:

$$Y = F\left(\mu + m_1(X^1) + \cdots + m_d(X^d)\right) + U, \tag{7}$$

$$E\left[U|\, X=x\right] = 0 \text{ for a.e. } x,$$

where F is a known link function. Again based on an i.i.d sample $\{Y_i, X_i\}_{i=1}^n$, we wish to conduct inference on the function $m_1(x^1)$ evaluated at some value $x^1\in[-1,1]$.

The model (7) is a natural generalization of the generalized linear model (see, e.g., McCullagh and Nelder, 1989) to the nonparametric context. Also note that this model is a generalization of the additive model (1), which corresponds to the case of $F(z) = z$. The model (7) is particularly useful to analyze the case where the response variable Y has a limited support. For example, if Y is binary (0 or 1), the nonparametric additive probit or logit model is specified by setting F as the normal or logistic cumulative distribution function, respectively. Also, if Y takes non-negative integers (i.e., count data), the nonparametric additive Poisson regression model is specified by setting $F(z) = \exp(z)$.

We extend the construction of the empirical likelihood function (6) in the last section to the generalized additive model. If we know the functions $m_2, \ldots, m_d$ and the intercept μ, then $m_1(x^1)$ can be estimated by the local (nonlinear) regression

$$\min_a \sum_{i=1}^n K_h\left(x^1 - X_i^1\right)\left\{Y_i - F\left(a + \mu + m_2(X_i^2) + \cdots + m_d(X_i^d)\right)\right\}^2, \tag{8}$$

where the solution $\hat{a}$ gives us an estimator of $m_1(x^1)$. Let $m_{-1}(\tilde{X}_i) = m_2(X_i^2) + \cdots + m_d(X_i^d)$ and $\tilde{X}_i = \left(X_i^2, \ldots, X_i^d\right)'$. By assuming that F is differentiable, the first-order condition of $\hat{a}$ is written as

$$\sum_{i=1}^n K_h\left(x^1 - X_i^1\right)\left\{Y_i - F\left(\hat{a} + \mu + m_{-1}(\tilde{X}_i)\right)\right\} F'\left(\hat{a} + \mu + m_{-1}(\tilde{X}_i)\right) = 0.$$

If we regard this condition as an estimating equation for $m_1(x^1)$, the empirical likelihood function for $m_1(x^1)$ can be defined as

$$L_F^*(a) = \sup_{\{p_i\}_{i=1}^n} \prod_{i=1}^n p_i,$$

$$\text{s.t. } 0 \le p_i \le 1, \ \sum_{i=1}^n p_i = 1, \ \sum_{i=1}^n p_i g_i(a) = 0,$$

where

$$g_i(a) = K_h\left(x^1 - X_i^1\right)\left\{Y_i - F\left(a + \mu + m_{-1}(\tilde{X}_i)\right)\right\} F'\left(a + \mu + m_{-1}(\tilde{X}_i)\right).$$

By applying the Lagrange multiplier method, the dual form for the empirical likelihood ratio $\ell_F^*(a) = -2\{\log L_F^*(a) + n \log n\}$ is obtained as

$$\ell_F^*(a) = 2 \sup_{\lambda \in \Lambda_{F,n}^*(a)} \sum_{i=1}^n \log(1 + \lambda g_i(a)),$$

where $\Lambda_{F,n}^*(a) = \{\lambda \in \mathbb{R} : \lambda g_i(a) \in \mathbb{V}_F^* \text{ for } i = 1, \ldots, n\}$ and $\mathbb{V}_F^*$ is an open interval containing 0. Again, since λ is scalar and the objective function $\sum_{i=1}^n \log(1 + \lambda g_i(a))$ is typically concave in λ, the computational cost to evaluate the empirical likelihood ratio $\ell_F^*(a)$ is not expensive.

Although we cannot compute $\ell_F^*(a)$ in practice, a feasible analog of $\ell_F^*(a)$ is available by replacing $\mu + m_{-1}(\tilde{X}_i)$ with its estimate. Similar to the case of the identity link function, we estimate $\mu + m_{-1}(\tilde{X}_i)$

based on a series approximation. By using the truncated basis functions $P_\kappa(x)$ defined in the last section, $\hat{\theta}_\kappa$ is defined as a solution to the least square problem:

$$\min_{\theta_\kappa \in \Theta_\kappa} \sum_{i=1}^{n} \left\{ Y_i - F\left(P_\kappa(X_i)' \theta_\kappa \right) \right\}^2,$$

where Θ_κ is a compact subset of $\mathbb{R}^{\kappa d+1}$ (due to the nonlinearity of the objective function, we need compactness of the parameter space). Note that μ is estimated by the first component of $\hat{\theta}_\kappa$ (denote by $\tilde{\mu}$) and $m_j(x^j)$ is estimated by an adequate component of $P_\kappa(x)' \hat{\theta}_\kappa$ (denote by $\tilde{m}_j(x^j)$). Then letting $\tilde{m}_{-1}(\tilde{X}_i) = \tilde{m}_2(X_i^2) + \cdots + \tilde{m}_d(X_i^d)$, an feasible analog of $\ell_F^*(a)$ is defined as

$$\ell_F(a) = 2 \sup_{\lambda \in \Lambda_{F,n}(a)} \sum_{i=1}^{n} \log\left(1 + \lambda \tilde{g}_i(a)\right), \tag{9}$$

where

$$\tilde{g}_i(a) = K_h\left(x^1 - X_i^1\right) \left\{ Y_i - F\left(a + \tilde{\mu} + \tilde{m}_{-1}(\tilde{X}_i) \right) \right\} F'\left(a + \tilde{\mu} + \tilde{m}_{-1}(\tilde{X}_i) \right),$$

$\Lambda_{F,n}(a) = \{\lambda \in \mathbb{R} : \lambda \tilde{g}_i(a) \in \mathbb{V}_F \text{ for } i = 1, \ldots, n\}$, and $\mathbb{V}_F$ is an open interval containing 0.

The asymptotic property of the empirical likelihood ratio $\ell_F(a)$ evaluated at $a = m_1(x^1)$ is obtained as follows.

Theorem 3.1. *Under Assumptions 1-5 in Appendix A.1,*

$$\ell_F\left(m_1(x^1)\right) \xrightarrow{d} \chi_1^2,$$

for each $x^1 \in [-1, 1]$.

The same remarks to Theorem 2.1 apply. In particular, the $100(1-\alpha)\%$ asymptotic empirical likelihood confidence interval for $m_1(x^1)$ is obtained as

$$ELCI_{F,\alpha} = \left\{ a : \ell_F(a) \leq \chi_{1,1-\alpha}^2 \right\}.$$

Remark 3.1 (Local linear fitting)**.** As in Section 2 for the identity link function case, we can also include the linear term of X^1 to the minimization problem in (8), i.e.,

$$\min_{a,b} \sum_{i=1}^{n} K_h\left(x^1 - X_i^1\right) \left\{ Y_i - F\left(a + b\left(X_i^1 - x^1\right) + \mu + m_{-1}(\tilde{X}_i) \right) \right\}^2.$$

However, in contrast to the identity link function case, the solution $\left(\hat{a}, \hat{b}\right)$ to the above minimization problem does not have an explicit form in general. Thus, to construct empirical likelihood, we need to incorporate the two-dimensional estimating equations:

$$\sum_{i=1}^{n} \left(\begin{array}{c} K_h\left(x^1 - X_i^1\right) \left\{ Y_i - F\left(\hat{a} + \hat{b}\left(X_i^1 - x^1\right) + \mu + m_{-1}(\tilde{X}_i) \right) \right\} \\ \times F'\left(\hat{a} + \hat{b}\left(X_i^1 - x^1\right) + \mu + m_{-1}(\tilde{X}_i) \right) \end{array} \right) \left[\begin{array}{c} 1 \\ X_i^1 - x^1 \end{array} \right] = 0.$$

Based these estimating equations, a feasible analog of the (profile) empirical likelihood ratio for $m_1(x^1)$ is defined as

$$\bar{\ell}_F(a) = \min_b \left\{ 2 \sup_{\lambda \in \bar{\Lambda}_{F,n}(a,b)} \sum_{i=1}^n \log\left(1 + \lambda' \bar{g}_i(a,b)\right) \right\}, \tag{10}$$

where

$$\bar{g}_i(a,b) = \begin{pmatrix} K_h\left(x^1 - X_i^1\right)\left\{Y_i - F\left(a + b\left(X_i^1 - x^1\right) + \tilde{\mu} + \tilde{m}_{-1}(\tilde{X}_i)\right)\right\} \\ \times F'\left(a + b\left(X_i^1 - x^1\right) + \tilde{\mu} + \tilde{m}_{-1}(\tilde{X}_i)\right) \end{pmatrix} \begin{bmatrix} 1 \\ X_i^1 - x^1 \end{bmatrix},$$

$\bar{\Lambda}_{F,n}(a,b) = \left\{\lambda \in \mathbb{R}^2 : \lambda' \bar{g}_i(a,b) \in \bar{\mathbb{V}}_F \text{ for } i = 1, \ldots, n\right\}$, and $\bar{\mathbb{V}}_F$ is an open interval containing 0. It should be noted that compared to the empirical likelihood ratio $\ell_F(a)$ in (9) based on local constant fitting, the empirical likelihood ratio $\bar{\ell}_F(a)$ based on local linear fitting requires additional minimization with respect to b and is computationally more expensive. In particular, to evaluate the empirical likelihood ratio $\bar{\ell}_F(a)$, we typically need to employ some nested algorithm (i.e., for each b we call a subroutine to implement optimization with respect to λ). This additional minimization step does not appear in the identity link function case because the estimating equations for $\left(\hat{a}, \hat{b}\right)$ can be solved explicitly. Although the technical argument will be more lengthy and complicated, we can expect that $\bar{\ell}_F\left(m_1(x^1)\right)$ converges in distribution to the χ_1^2 distribution as well as $\ell_F\left(m_1(x^1)\right)$.

4 Additive Quantile Regression

We finally consider the nonparametric additive quantile regression model:

$$Y = \mu + m_1(X^1) + \cdots + m_d(X^d) + U, \tag{11}$$

$$Q_\tau\left(U \middle| X = x\right) = 0 \text{ for a.e. } x,$$

where $Q_\tau\left(\cdot \middle| X = x\right)$ denotes the τ-th conditional quantile function given $X = x$ with $\tau \in (0,1)$. A special case is the additive median regression with $\tau = 0.5$. This model, studied by e.g., Doksum and Koo (2000), Goojier and Zerom (2003), and Horowitz and Lee (2005), is a generalization of the additive model (1) for the conditional mean to the conditional quantiles. Based on an i.i.d sample $\{Y_i, X_i\}_{i=1}^n$, we wish to conduct inference on the function $m_1(x^1)$ evaluated at some value $x^1 \in [-1, 1]$.

The construction of the empirical likelihood function $\ell(a)$ in (6) for the conditional mean case can be extended as follows. If we know the functions $m_2, \ldots, m_d$ and the intercept μ, then let $Y_i^* = Y_i - \mu - m_2(X_i^2) - \cdots - m_d(X_i^d)$ again and we can estimate $m_1(x^1)$ by the local (constant) quantile regression

$$\min_a \sum_{i=1}^n K_h\left(x^1 - X_i^1\right) \rho_\tau\left(Y_i^* - a\right), \tag{12}$$

where $\rho_\tau(v) = v\left(\tau - I\{v \leq 0\}\right)$ is the so-called check function (Koenker and Bassett, 1978) and $I\{\cdot\}$ is the indicator function. The solution $\hat{a}$ gives us an estimator of $m_1(x^1)$. By taking the derivative except for the point with $Y_i^* - a = 0$, the (asymptotic) first-order condition of $\hat{a}$ is written as

$$\sum_{i=1}^{n} K_h \left(x^1 - X_i^1\right) \left(\tau - I\left\{Y_i^* \leq a\right\}\right) = 0.$$

If we regard this condition as an estimating equation for $m_1(x^1)$, the empirical likelihood function for $m_1(x^1)$ can be defined as

$$L_\tau^*(a) = \sup_{\{p_i\}_{i=1}^n} \prod_{i=1}^{n} p_i,$$

$$\text{s.t. } 0 \leq p_i \leq 1, \ \sum_{i=1}^{n} p_i = 1, \ \sum_{i=1}^{n} p_i K_h \left(x^1 - X_i^1\right) \left(\tau - I\left\{Y_i^* \leq a\right\}\right) = 0.$$

By applying the Lagrange multiplier method, the dual form for the empirical likelihood ratio $\ell_\tau^*(a) = -2\left\{\log L_\tau^*(a) + n \log n\right\}$ is obtained as

$$\ell_\tau^*(a) = 2 \sup_{\lambda \in \Lambda_{\tau,n}^*(a)} \sum_{i=1}^{n} \log\left(1 + \lambda K_h \left(x^1 - X_i^1\right) \left(\tau - I\left\{Y_i^* \leq a\right\}\right)\right),$$

where $\Lambda_{\tau,n}^*(a) = \left\{\lambda \in \mathbb{R} : \lambda K_h \left(x^1 - X_i^1\right) \left(\tau - I\left\{Y_i^* \leq a\right\}\right) \in \mathbb{V}_\tau^* \text{ for } i = 1, \ldots, n\right\}$ and $\mathbb{V}_\tau^*$ is an open interval containing 0.

Note that although the objective function $\sum_{i=1}^n \log\left(1 + \lambda K_h \left(x^1 - X_i^1\right) \left(\tau - I\left\{Y_i^* \leq a\right\}\right)\right)$ is non-smooth in a, it is smooth in λ. Therefore, we can still apply the conventional Newton-type gradient-based optimization to evaluate $\ell_\tau^*(a)$.

Similar to the previous sections, a feasible analog of $\ell_\tau^*(a)$ is obtained by replacing $\mu + m_2(X_i^2) + \cdots + m_d(X_i^d)$ with its estimate. By using the truncated basis functions $P_\kappa(x)$ defined in Section 2, $\hat{\theta}_\kappa$ is defined as a solution to the quantile regression problem:

$$\min_{\theta_\kappa} \sum_{i=1}^{n} \rho_\tau \left(Y_i - P_\kappa(X_i)' \theta_\kappa\right).$$

Since this is the conventional linear quantile regression problem, we can apply the standard algorithm such as the linear programming method (see, e.g., Koenker, 2005). Note that μ is estimated by the first component of $\hat{\theta}_\kappa$ (denote by $\tilde{\mu}$) and $m_j(x^j)$ is estimated by an adequate component of $P_\kappa(x)' \hat{\theta}_\kappa$ (denote by $\tilde{m}_j(x^j)$). Then letting $\tilde{Y}_i = Y_i - \tilde{\mu} - \tilde{m}_2(X_i^2) - \cdots - \tilde{m}_d(X_i^d)$, an feasible analog of $\ell_\tau^*(a)$ is defined as

$$\ell_\tau(a) = 2 \sup_{\lambda \in \Lambda_{\tau,n}(a)} \sum_{i=1}^{n} \log\left(1 + \lambda K_h \left(x^1 - X_i^1\right) \left(\tau - I\left\{\tilde{Y}_i \leq a\right\}\right)\right),$$

where $\Lambda_{\tau,n}(a) = \left\{\lambda \in \mathbb{R} : \lambda K_h \left(x^1 - X_i^1\right) \left(\tau - I\left\{\tilde{Y}_i \leq a\right\}\right) \in \mathbb{V}_\tau \text{ for } i = 1, \ldots, n\right\}$, and $\mathbb{V}_\tau$ is an open interval containing 0.

The asymptotic property of the empirical likelihood ratio $\ell_\tau(a)$ evaluated at $a = m_1(x^1)$ is obtained as follows.

Theorem 4.1. *Under Assumptions 1-4 and 6 in Appendix A.1,*

$$\ell_\tau\left(m_1(x^1)\right) \xrightarrow{d} \chi_1^2,$$

for each $x^1 \in [-1,1]$ and $\tau \in (0,1)$.

The same remarks to Theorem 2.1 apply. In particular, the $100\,(1-\alpha)\,\%$ asymptotic empirical likelihood confidence interval for $m_1(x^1)$ is obtained as

$$ELCI_{\tau,\alpha} = \left\{a : \ell_\tau\left(a\right) \leq \chi^2_{1,1-\alpha}\right\}.$$

Remark 4.1 (Local linear fitting)**.** As in Section 2 for the identity link function case, we can include the linear term of X^1 to the minimization problem in (12), i.e.,

$$\min_{a,b}\sum_{i=1}^n K_h\left(x^1 - X_i^1\right)\rho_\tau\left(Y_i^* - a - b\left(X_i^1 - x^1\right)\right).$$

However, similar to the non-identity link function case in Section 3, the solution $\left(\hat{a},\hat{b}\right)$ to the above minimization problem does not have an explicit form in general. Thus, to construct empirical likelihood, we need to incorporate the two-dimensional estimating equations:

$$\sum_{i=1}^n K_h\left(x^1 - X_i^1\right)\left(\tau - I\left\{Y_i^* \leq \hat{a} + \hat{b}\left(X_i^1 - x^1\right)\right\}\right)\left[\begin{array}{c} 1 \\ X_i^1 - x^1 \end{array}\right] = 0.$$

Based these estimating equations, a feasible analog of the (profile) empirical likelihood ratio for $m_1(x^1)$ is defined as

$$\bar{\ell}_\tau\left(a\right) = \min_b\left\{2\sup_{\lambda\in\bar{\Lambda}_{\tau,n}(a,b)}\sum_{i=1}^n \log\left(1 + \lambda'\bar{g}_{\tau,i}\left(a,b\right)\right)\right\}, \tag{13}$$

where

$$\bar{g}_{\tau,i}\left(a,b\right) = K_h\left(x^1 - X_i^1\right)\left(\tau - I\left\{\tilde{Y}_i \leq a + b\left(X_i^1 - x^1\right)\right\}\right)\left[\begin{array}{c} 1 \\ X_i^1 - x^1 \end{array}\right],$$

$\bar{\Lambda}_{\tau,n}\left(a,b\right) = \left\{\lambda \in \mathbb{R}^2 : \lambda'\bar{g}_{\tau,i}\left(a,b\right) \in \bar{\mathbb{V}}_\tau \text{ for } i = 1,\ldots,n\right\}$, and $\bar{\mathbb{V}}_\tau$ is an open interval containing 0. Similar to (10), the empirical likelihood ratio $\bar{\ell}_\tau\left(a\right)$ based on local linear fitting requires additional minimization with respect to b. Note that this minimization is computationally more demanding than the one in (10) because the objective function for b is generally non-smooth. Therefore, although we can expect that $\bar{\ell}_\tau\left(m_1(x^1)\right)$ converges in distribution to the χ_1^2 distribution by more elaborate technical arguments, we do not recommend this approach due to the practical drawback.[4]

[4] As in Otsu (2008), it is possible to replace the indicator function in $\bar{g}_{\tau,i}\left(a,b\right)$ with an integrated kernel function to make the objective function for b smooth.

5 Conclusion

This paper proposes empirical likelihood inference methods for three types of nonparametric additive models: additive mean regression with the identity link function, generalized additive mean regression with a known non-identity link function, and additive quantile regression. For these models, we construct empirical likelihood functions and derive the empirical likelihood ratio statistics for the unknown functions. The associated confidence intervals obtained from inverting the empirical likelihood ratio statistics have attractive features compared to the conventional Wald-type confidence intervals. It is interesting to extend the present approach to other nonparametric settings, such as additive regression with an unknown link function and censored additive regression.

A Mathematical Appendix

A.1 Assumptions

1. (Assumptions on data)

   (i) For almost every $x \in [-1,1]^d$,

$$E\left[Y | X = x\right] = \begin{cases} \mu + m_1(x^1) + \cdots + m_d(x^d) & \text{(identity link case)} \\ F\left(\mu + m_1(x^1) + \cdots + m_d(x^d)\right) & \text{(non-identity link case)} \end{cases}$$

$$Q_\tau\left(Y | X = x\right) = \mu + m_1(x^1) + \cdots + m_d(x^d) \qquad \text{(quantile case)}$$

   (ii) $\{Y_i, X_i\}_{i=1}^n$ is an i.i.d. sample of (Y, X).

   (iii) X is absolutely continuous with respect to the Lebesgue measure with the support $X \in [-1,1]^d$.

   (iv) The density function of X is bounded, bounded away from zero, twice continuously differentiable in the interior of $[-1,1]^d$, and has continuous second-order one-sided derivatives at the boundary of $[-1,1]^d$.

   (v) $Var\left(U | X = x\right)$ is bounded and bounded away from zero for all $x \in [-1,1]^d$.

   (vi) $E\left|U\right|^j \leq C^{j-2} j! E\left[U^2\right]$ for all $j \geq 2$ and some $C \in (0, \infty)$.

2. (Assumptions on m_j)

   (i) $|m_j(v)| \leq C_m < \infty$ for all $v \in [-1,1]$ and all $j = 1, \ldots, d$.

   (ii) m_j is twice continuously differentiable in the interior of $[-1,1]$ and has continuous second-order one-sided derivatives at the boundary of $[-1,1]$ for all $j = 1, \ldots, d$.

3. (Assumptions on basis and series length)

   (i) $\{p_k\}$ satisfies $\int_{-1}^{1} p_j\left(v\right) p_k\left(v\right) dv = I\left\{j = k\right\}$ and $\int_{-1}^{1} p_k\left(v\right) dv = 0$ for all $j, k \in \mathbb{N}$.

   (ii) $\sup_{x \in [-1,1]^d} \left|P_\kappa\left(x\right)\right|$ is bounded away from zero for all κ large enough and $\sup_{x \in [-1,1]^d} \left|P_\kappa\left(x\right)\right| = O\left(\kappa^{1/2}\right)$ as $\kappa \to \infty$.

   (iii) There exists $\theta_{\kappa 0} \in \mathbb{R}^{d\kappa+1}$ (identity link and quantile cases) or $\theta_{\kappa 0} \in \Theta_\kappa$ (non-identity link case) such that $\sup_{x \in [-1,1]^d} \left|\mu + m_1(x^1) + \cdots + m_d(x^d) - P_\kappa\left(x\right)' \theta_{\kappa 0}\right| = O\left(\kappa^{-2}\right)$ as $\kappa \to \infty$.

   (iv) The smallest eigenvalue of

$$Q_\kappa = \begin{cases} E\left[P_\kappa\left(X\right) P_\kappa\left(X\right)'\right] & \text{(identity link case)} \\ E\left[F'\left(\mu + m_1(X^1) + \cdots + m_d(X^d)\right)^2 P_\kappa\left(X\right) P_\kappa\left(X\right)'\right] & \text{(non-identity link case)} \\ E\left[f_U\left(0 | X\right) P_\kappa\left(X\right) P_\kappa\left(X\right)'\right] & \text{(quantile case)} \end{cases}$$

is bounded away from zero for all $\kappa \in \mathbb{N}$, where $f_U(\cdot|\, x)$ is the conditional density function of U in (11) given $X = x$. Each element of Q_κ is bounded for all $\kappa \in \mathbb{N}$.

(v) The largest eigenvalue of

$$\Psi_\kappa = \begin{cases} Q_\kappa^{-1} E\left[Var\left(U|\, X\right) P_\kappa\left(X\right) P_\kappa\left(X\right)'\right] Q_\kappa^{-1} & \text{(identity link case)} \\ Q_\kappa^{-1} E\left[\begin{array}{c} F'\left(\mu + m_1(X^1) + \cdots + m_d(X^d)\right)^2 \\ \times Var\left(U|\, X\right) P_\kappa\left(X\right) P_\kappa\left(X\right)' \end{array}\right] Q_\kappa^{-1} & \text{(non-identity link case)} \end{cases}$$

is bounded for all $\kappa \in \mathbb{N}$. For the quantile case, let

$$\bar{P}_\kappa\left(\tilde{x}\right) = \left[1, \underbrace{0, \ldots, 0}_{k}, p_1(x^2), \ldots, p_\kappa(x^2), \ldots, p_1(x^d), \ldots, p_\kappa(x^d)\right]',$$

and the largest eigenvalue of $\Psi_k\left(x^1\right) = E\left[\bar{P}_\kappa\left(\tilde{X}\right)\bar{P}_\kappa\left(\tilde{X}\right)'\middle|\, X^1 = x^1\right]$ is bounded for all $\kappa \in \mathbb{N}$ and $x^1 \in [-1, 1]$, and twice continuously differentiable in the interior of $[-1, 1]$ for all $\kappa \in \mathbb{N}$, and has continuous second-order one-sided derivatives at the boundary of $[-1, 1]$ for all $\kappa \in \mathbb{N}$.

(vi) $\kappa = C_\kappa n^v$ for some $C_\kappa \in (0, \infty)$ and some $v \in \left(\frac{4}{15}, \frac{3}{10}\right)$ (identity and non-identity link cases) or $v \in \left(\frac{1}{5}, \frac{7}{30}\right)$ (quantile case).

4. (Assumptions on kernel and bandwidth)

   (i) K is a bounded, continuous, and symmetric (around zero) density function on $[-1, 1]$.

   (ii) As $n \to \infty$, it holds $h \to 0$, $nh \to \infty$, and $nh^5 \to 0$.

5. (Additional assumptions for non-identity link function)

   (i) For all $v \in [\mu - C_m d, \mu + C_m d]$, $F(v)$ is bounded, F is twice continuously differentiable, and $F'(v)$ is bounded and bounded away from zero. There exists a constant $C_F \in (0, \infty)$ such that $|F''(v_1) - F''(v_2)| \leq C_F |v_1 - v_2|$ for all $v_1, v_2 \in [\mu - C_m d, \mu + C_m d]$.

   (ii) For some constant $C_\theta \in (0, \infty)$, $\Theta_\kappa = [-C_\theta, C_\theta]^{\kappa d + 1}$ for all $\kappa \in \mathbb{N}$. For all $\kappa \in \mathbb{N}$, θ_κ is in interior of Θ_κ.

6. (Additional assumptions for quantile regression)

   The conditional distribution function $F_U(u|\, x)$ of U in (11) given $X = x$ satisfies $F_U(0|\, x) = \tau$ for almost every $x \in [-1, 1]^d$, and has a density function $f_U(u|\, x)$ which is bounded and bounded away from zero for all u in a neighborhood of 0 and for all $x \in [-1, 1]^d$. There exists a constant $C_f \in (0, \infty)$ such that $|f_U(u_1|\, x) - f_U(u_2|\, x)| \leq C_f |u_1 - u_2|$ for all u_1 and u_2 in a neighborhood of 0 and for all $x \in [-1, 1]^d$

A.2 Proof of Theorem 2.1

In this subsection, let f_1 be the density function of X^1, and

$$\begin{aligned} S_{n,j} &= \frac{1}{nh}\sum_{j=1}^{n} K_h\left(x^1 - X_i^1\right)\left(\frac{X_i^1 - x^1}{h}\right)^j, \qquad s_{j_1 j_2} = f_1\left(x^1\right)\int K\left(z\right)^{j_1} z^{j_2} dz, \\ \sigma^2 &= Var\left(U \middle| X^1 = x^1\right), \qquad V = \sigma^2 s_{12}^2 s_{20}, \\ g_i &= \tilde{K}_i\left(\tilde{Y}_i - m_1(x^1)\right). \end{aligned}$$

From Lemma A.1 (iii), the first-order condition for $\hat{\lambda}$ satisfies

$$0 = \frac{1}{nh}\sum_{i=1}^{n}\frac{g_i}{1+\hat{\lambda}g_i} = \frac{1}{nh}\sum_{i=1}^{n} g_i - \hat{V}_1\hat{\lambda},$$

w.p.a.1 (with probability approaching one), where $\hat{V}_1 = \frac{1}{nh}\sum_{i=1}^{n}\frac{g_i^2}{\left(1+\dot{\lambda}g_i\right)^2}$, the second equality follows from an expansion around $\hat{\lambda} = 0$, and $\dot{\lambda}$ is a point on the line joining $\hat{\lambda}$ and 0. Since

$$\left|\hat{V}_1 - V\right| \leq \max_{1\leq i\leq n}\left|\frac{1}{1+\dot{\lambda}g_i}\right|^2\left|\frac{1}{nh}\sum_{i=1}^{n} g_i^2 - V\right| \xrightarrow{p} 0,$$

(by Lemma A.1 (ii) and (iii)) and $V > 0$, $\hat{V}_1^{-1}$ exists w.p.a.1. Thus, we have

$$\hat{\lambda} = \hat{V}_1^{-1}\frac{1}{nh}\sum_{i=1}^{n} g_i,$$

w.p.a.1. From Lemma A.1 (iii), $\hat{\lambda}$ satisfies $\ell\left(m_1(x^1)\right) = 2\sum_{i=1}^{n}\log\left(1+\hat{\lambda}g_i\right)$ w.p.a.1, and a second-order expansion of this equation around $\hat{\lambda} = 0$ yields

$$\ell\left(m_1(x^1)\right) = 2\hat{\lambda}\sum_{i=1}^{n} g_i - \hat{V}_2\hat{\lambda}^2 = \left[2\hat{V}_1^{-1} - \hat{V}_2\hat{V}_1^{-2}\right]\left(\frac{1}{\sqrt{nh}}\sum_{i=1}^{n} g_i\right)^2,$$

w.p.a.1, where $\hat{V}_2 = \frac{1}{nh}\sum_{i=1}^{n}\frac{g_i^2}{\left(1+\ddot{\lambda}g_i\right)^2}$ and $\ddot{\lambda}$ is a point on the line joining $\hat{\lambda}$ and 0. Since $\left|\hat{V}_2 - V\right| \xrightarrow{p} 0$ by the same argument to $\hat{V}_1$, we have $2\hat{V}_1^{-1} - \hat{V}_2\hat{V}_1^{-2} \xrightarrow{p} V^{-1}$. Therefore, Lemma A.1 (ii) implies the conclusion.

Lemma A.1. *Under Assumptions 1-4 in Appendix A.1, it holds*

(i) $S_{n,1} = O_p\left((nh)^{-1/2}\right) + O\left(h^2\right)$, *and* $S_{n,2} - s_{12} = O_p\left((nh)^{-1/2}\right) + O\left(h^2\right)$;

(ii) $\frac{1}{nh}\sum_{i=1}^{n} g_i^2 \xrightarrow{p} V$ *and* $\frac{1}{\sqrt{nh}}\sum_{i=1}^{n} g_i \xrightarrow{d} N\left(0, V\right)$;

(iii) *there exists* $\hat{\lambda} \in int\left(\Lambda_n\left(m_1(x^1)\right)\right)$ *satisfying*

$$\sum_{i=1}^{n}\log\left(1+\hat{\lambda}g_i\right) = \sup_{\lambda\in\Lambda_n(m_1(x^1))}\sum_{i=1}^{n}\log\left(1+\lambda g_i\right), \quad \textit{w.p.a.1},$$

$\left|\hat{\lambda}\right| = O_p\left((nh)^{-1/2}\right)$, *and* $\max_{1\leq i\leq n}\left|\hat{\lambda}g_i\right| \xrightarrow{p} 0$.

Proof of (i). We only prove the first statement. The second statement can be shown in the same manner. By the change of variables and a second-order expansion of $f_1\left(x^1+hz\right)$ around $hz=0$, we have

$$E\left[S_{n,1}\right]=\int K\left(z\right)zf_1\left(x^1+hz\right)dz=O\left(h^2\right).$$

Also, a similar argument yields

$$Var\left(S_{n,1}\right)\leq\frac{1}{nh^2}E\left[K_h\left(x^1-X_i^1\right)^2\left(\frac{X_i^1-x^1}{h}\right)^2\right]=\frac{1}{nh}\int K\left(z\right)^2z^2f_1\left(x^1+hz\right)dz=O\left(\left(nh\right)^{-1}\right).$$

Therefore, Lyapunov's central limit theorem implies $S_{n,1}-E\left[S_{n,1}\right]=O_p\left(\left(nh\right)^{-1/2}\right)$. Combining these results, the conclusion is obtained.

Proof of (ii). Proof of the first statement. From $\tilde{K}_i=K_h\left(x^1-X_i^1\right)\left\{S_{n,2}-\left(\frac{X_i^1-x^1}{h}\right)S_{n,1}\right\}$,

$$\begin{aligned}\frac{1}{nh}\sum_{i=1}^n g_i^2 &= \frac{1}{nh}\sum_{i=1}^n\tilde{K}_i^2\left(\tilde{Y}_i-m_1(x^1)\right)^2\\ &= S_{n,2}^2\frac{1}{nh}\sum_{i=1}^n K_h\left(x^1-X_i^1\right)^2\left(\tilde{Y}_i-m_1(x^1)\right)^2\\ &\quad+S_{n,1}^2\frac{1}{nh}\sum_{i=1}^n K_h\left(x^1-X_i^1\right)^2\left(\frac{X_i^1-x^1}{h}\right)^2\left(\tilde{Y}_i-m_1(x^1)\right)^2\\ &\quad-2S_{n,2}S_{n,1}\frac{1}{nh}\sum_{i=1}^n K_h\left(x^1-X_i^1\right)^2\left(\frac{X_i^1-x^1}{h}\right)\left(\tilde{Y}_i-m_1(x^1)\right)^2\\ &= T_1+T_2-2T_3.\end{aligned}$$

For T_1, note that

$$\begin{aligned}T_1 &= S_{n,2}^2\frac{1}{nh}\sum_{i=1}^n K_h\left(x^1-X_i^1\right)^2\left(Y_i^*-m_1(x^1)\right)^2+S_{n,2}^2\frac{1}{nh}\sum_{i=1}^n K_h\left(x^1-X_i^1\right)^2\left(\tilde{Y}_i-Y_i^*\right)^2\\ &\quad+2S_{n,2}^2\frac{1}{nh}\sum_{i=1}^n K_h\left(x^1-X_i^1\right)^2\left(Y_i^*-m_1(x^1)\right)\left(\tilde{Y}_i-Y_i^*\right)\\ &= T_{11}+T_{12}+2T_{13}.\end{aligned}$$

By the same argument to the proof of Part (i) of this lemma,

$$\begin{aligned}E\left[\frac{1}{nh}\sum_{i=1}^n K_h\left(x^1-X_i^1\right)^2\left(Y_i^*-m_1(x^1)\right)^2\right] &\to \sigma^2 s_{20},\\ Var\left(\frac{1}{nh}\sum_{i=1}^n K_h\left(x^1-X_i^1\right)^2\left(Y_i^*-m_1(x^1)\right)^2\right) &\to 0,\end{aligned}\tag{14}$$

Thus, from Chebyshev's inequality and Lemma A.1 (i), we have $T_{11}\xrightarrow{p}\sigma^2s_{20}$. For T_{12} and T_{13}, adapted versions of Horowitz and Mammen (2004, Lemma 7), where the link function is set to identity, and Lemma A.1 (i) imply that $T_{12}\xrightarrow{p}0$ and $T_{13}\xrightarrow{p}0$. Combining these results, the conclusion is obtained.

Proof of the second statement. Again, from $\tilde{K}_i = K_h\left(x^1 - X_i^1\right)\left\{S_{n,2} - \left(\frac{X_i^1 - x^1}{h}\right)S_{n,1}\right\}$,

$$\begin{aligned}
\frac{1}{\sqrt{nh}}\sum_{i=1}^{n} g_i &= \frac{1}{\sqrt{nh}}\sum_{i=1}^{n}\tilde{K}_i\left(\tilde{Y}_i - m_1(x^1)\right) \\
&= \frac{1}{\sqrt{nh}}\sum_{i=1}^{n} K_h\left(x^1 - X_i^1\right)\left\{S_{n,2} - \left(\frac{X_i^1 - x^1}{h}\right)S_{n,1}\right\}\left(Y_i^* - m_1(x^1)\right) \\
&\quad + \frac{1}{\sqrt{nh}}\sum_{i=1}^{n} K_h\left(x^1 - X_i^1\right)\left\{S_{n,2} - \left(\frac{X_i^1 - x^1}{h}\right)S_{n,1}\right\}\left(\tilde{Y}_i - Y_i^*\right) \\
&= L_1 + L_2.
\end{aligned}$$

For L_1, note that

$$\begin{aligned}
L_1 &= \left(S_{n,2} - s_{12}\right)\frac{1}{\sqrt{nh}}\sum_{i=1}^{n} K_h\left(x^1 - X_i^1\right)\left(Y_i^* - m_1(x^1)\right) \\
&\quad - \left(S_{n,1} - s_{11}\right)\frac{1}{\sqrt{nh}}\sum_{i=1}^{n} K_h\left(x^1 - X_i^1\right)\left(\frac{X_i^1 - x^1}{h}\right)\left(Y_i^* - m_1(x^1)\right) \\
&\quad + \frac{1}{\sqrt{nh}}\sum_{i=1}^{n}\left\{\begin{array}{c} K_h\left(x^1 - X_i^1\right)\left\{s_{12} - \left(\frac{X_i^1 - x^1}{h}\right)s_{11}\right\}\left(Y_i^* - m_1(x^1)\right) \\ -E\left[K_h\left(x^1 - X_i^1\right)\left\{s_{12} - \left(\frac{X_i^1 - x^1}{h}\right)s_{11}\right\}\left(Y_i^* - m_1(x^1)\right)\right] \end{array}\right\} \\
&\quad + \sqrt{\frac{n}{h}}E\left[K_h\left(x^1 - X_i^1\right)\left\{s_{12} - \left(\frac{X_i^1 - x^1}{h}\right)s_{11}\right\}\left(Y_i^* - m_1(x^1)\right)\right] \\
&= L_{11} - L_{12} + L_{13} + L_{14}.
\end{aligned}$$

For L_{11}, Lyapunov's central limit theorem implies

$$\frac{1}{\sqrt{nh}}\sum_{i=1}^{n}\left\{K_h\left(x^1 - X_i^1\right)\left(Y_i^* - m_1(x^1)\right) - E\left[K_h\left(x^1 - X_i^1\right)\left(Y_i^* - m_1(x^1)\right)\right]\right\} \xrightarrow{d} N\left(0, \sigma^2 s_{20}\right),$$

and the change of variables and a second-order expansion of $E\left[Y_i^* - m_1(x^1)\middle| X^1 = x^1 + hz\right] f_1\left(x^1 + hz\right)$ around $hz = 0$ imply

$$E\left[K_h\left(x^1 - X_i^1\right)\left(Y_i^* - m_1(x^1)\right)\right] = h\int K(z)\,E\left[Y_i^* - m_1(x^1)\middle| X^1 = x^1 + hz\right] f_1\left(x^1 + hz\right)dz = O\left(h^3\right).$$

Thus, from Lemma A.1 (i) and $nh^5 \to 0$, we have $L_{11} \xrightarrow{p} 0$. Similarly, we can show that $L_{12} \xrightarrow{p} 0$. For L_{13}, note that

$$\begin{aligned}
E\left[L_{13}\right] &= \int K(z)^2\left(s_{12} - s_{11}z\right)^2 E\left[\left(Y_i^* - m_1(x^1)\right)^2\middle| X^1 = x^1 + hz\right] f_1\left(x^1 + hz\right)dz \\
&\quad - h\left(\int K(z)\left(s_{12} - s_{11}z\right)E\left[Y_i^* - m_1(x^1)\middle| X^1 = x^1 + hz\right] f_1\left(x^1 + hz\right)dz\right)^2 \\
&\to V,
\end{aligned}$$

where the convergence follows from a similar argument to (14). Therefore, Lyapunov's central limit theorem implies $L_{13} \xrightarrow{d} N(0, V)$. For L_{14}, the change of variables and second-order expansion of

$E\left[\left. Y_i^* - m_1(x^1)\right| X^1 = x^1 + hz\right] f_1\left(x^1 + hz\right)$ around $hz = 0$ yield

$$L_{14} \;=\; \sqrt{nh}\int K\left(z\right)\left(s_{12} - s_{11}z\right) E\left[\left. Y_i^* - m_1(x^1)\right| X^1 = x^1 + hz\right] f_1\left(x^1 + hz\right) dz \to 0.$$

Combining these results, we obtain $L_1 \overset{d}{\to} N\left(0, V\right)$. On the other hand, from Horowitz and Mammen (2004, Lemma 7) with the identity link function and Lemma A.1 (i), we have $L_2 \overset{p}{\to} 0$. Therefore, the conclusion is obtained.

Proof of (iii). Since the proof is similar to Newey and Smith (2004, Lemmas A1 and A2), it is omitted.

A.3 Proof of Theorem 3.1

In this subsection, let f be the density function of X, $g_{F,i} = \tilde{g}_i\left(m_1(x^1)\right)$, and

$$V_F = \left(\int K\left(z\right)^2 dz\right)\int Var\left(\left. U\right| X = \left(x^1, \tilde{x}\right)'\right) F'\left(\mu + m_1(x^1) + m_{-1}(\tilde{x})\right)^2 f\left(x^1, \tilde{x}\right) d\tilde{x}.$$

From Lemma A.2 (ii), the first-order condition for $\hat{\lambda}_F$ satisfies

$$0 = \frac{1}{nh}\sum_{i=1}^n \frac{g_{F,i}}{1 + \hat{\lambda}_F g_{F,i}} = \frac{1}{nh}\sum_{i=1}^n g_{F,i} - \hat{V}_{F,1}\hat{\lambda}_F,$$

w.p.a.1, where $\hat{V}_{F,1} = \frac{1}{nh}\sum_{i=1}^n \frac{g_{F,i}^2}{\left(1+\dot{\lambda}_F g_{F,i}\right)^2}$, the second equality follows from an expansion around $\hat{\lambda}_F = 0$, and $\dot{\lambda}_F$ is a point on the line joining $\hat{\lambda}_F$ and 0. Since

$$\left|\hat{V}_{F,1} - V_F\right| \leq \max_{1\leq i\leq n}\left|\frac{1}{1 + \dot{\lambda}_F g_{F,i}}\right|^2 \left|\frac{1}{nh}\sum_{i=1}^n g_{F,i}^2 - V_F\right| \overset{p}{\to} 0,$$

(by Lemma A.2 (i) and (ii)) and $V_F > 0$ (by Assumptions 1 (iv)-(v), 4 (i), and 5 (v)), $\hat{V}_{F,1}^{-1}$ exists w.p.a.1. Thus, we have

$$\hat{\lambda}_F = \hat{V}_{F,1}^{-1}\frac{1}{nh}\sum_{i=1}^n g_{F,i},$$

w.p.a.1. From Lemma A.2 (ii), $\hat{\lambda}_F$ satisfies $\ell_F\left(m_1(x^1)\right) = 2\sum_{i=1}^n \log\left(1 + \hat{\lambda}_F g_{F,i}\right)$ w.p.a.1, and a second-order expansion of this equation around $\hat{\lambda}_F = 0$ yields

$$\ell_F\left(m_1(x^1)\right) = 2\hat{\lambda}_F\sum_{i=1}^n g_{F,i} - \hat{V}_{F,2}\hat{\lambda}_F^2 = \left[2\hat{V}_{F,1}^{-1} - \hat{V}_{F,2}\hat{V}_{F,1}^{-2}\right]\left(\frac{1}{\sqrt{nh}}\sum_{i=1}^n g_{F,i}\right)^2,$$

w.p.a.1, where $\hat{V}_{F,2} = \frac{1}{nh}\sum_{i=1}^n \frac{g_{F,i}^2}{\left(1+\ddot{\lambda}_F g_{F,i}\right)^2}$ and $\ddot{\lambda}_F$ is a point on the line joining $\hat{\lambda}_F$ and 0. Since $\left|\hat{V}_{F,2} - V_F\right| \overset{p}{\to} 0$ by the same argument to $\hat{V}_{F,1}$, we have $2\hat{V}_{F,1}^{-1} - \hat{V}_{F,2}\hat{V}_{F,1}^{-2} \overset{p}{\to} V_F^{-1}$. Therefore, Lemma A.2 (i) implies the conclusion.

Lemma A.2. *Under Assumptions 1-5 in Appendix A.1, it holds*

(i) $\frac{1}{nh}\sum_{i=1}^{n} g_{F,i}^2 \xrightarrow{p} V_F$ *and* $\frac{1}{\sqrt{nh}}\sum_{i=1}^{n} g_{F,i} \xrightarrow{d} N\left(0, V_F\right)$*;*

(ii) *there exists* $\hat{\lambda}_F \in int\left(\Lambda_{F,n}\left(m_1(x^1)\right)\right)$ *satisfying*

$$\sum_{i=1}^{n} \log\left(1+\hat{\lambda}_F g_{F,i}\right) = \sup_{\lambda\in\Lambda_{F,n}(m_1(x^1))} \sum_{i=1}^{n} \log\left(1+\lambda g_{F,i}\right), \quad w.p.a.1,$$

$\left|\hat{\lambda}_F\right| = O_p\left((nh)^{-1/2}\right)$*, and* $\max_{1\le i\le n}\left|\hat{\lambda}_F g_{F,i}\right| \xrightarrow{p} 0$.

Proof of (i). Proof of the first statement. Let

$$M_i = \mu + m_1(x^1) + m_{-1}(\tilde{X}_i), \qquad \tilde{M}_i = \tilde{\mu} + m_1(x^1) + \tilde{m}_{-1}(\tilde{X}_i).$$

By the definition of $g_{F,i}$ and expansions around $\tilde{M}_i = M_i$,

$$\begin{aligned}
\frac{1}{nh}\sum_{i=1}^{n} g_{F,i}^2 &= \frac{1}{nh}\sum_{i=1}^{n} K_h\left(x^1 - X_i^1\right)^2 \left\{U_i - F'\left(\dot{M}_i\right)\left(\tilde{M}_i - M_i\right)\right\}^2 \left\{F'\left(M_i\right) + F''\left(\ddot{M}_i\right)\left(\tilde{M}_i - M_i\right)\right\}^2 \\
&= \frac{1}{nh}\sum_{i=1}^{n} K_h\left(x^1 - X_i^1\right)^2 U_i^2 F'\left(M_i\right)^2 \\
&\quad + \frac{1}{nh}\sum_{i=1}^{n} K_h\left(x^1 - X_i^1\right)^2 F'\left(\dot{M}_i\right)^2 F'\left(M_i\right)^2 \left(\tilde{M}_i - M_i\right)^2 \\
&\quad - \frac{2}{nh}\sum_{i=1}^{n} K_h\left(x^1 - X_i^1\right)^2 U_i F'\left(\dot{M}_i\right) F'\left(M_i\right)^2 \left(\tilde{M}_i - M_i\right) \\
&\quad + \frac{1}{nh}\sum_{i=1}^{n} K_h\left(x^1 - X_i^1\right)^2 U_i^2 F''\left(\ddot{M}_i\right)^2 \left(\tilde{M}_i - M_i\right)^2 \\
&\quad + \frac{1}{nh}\sum_{i=1}^{n} K_h\left(x^1 - X_i^1\right)^2 F'\left(\dot{M}_i\right)^2 F''\left(\ddot{M}_i\right)^2 \left(\tilde{M}_i - M_i\right)^4 \\
&\quad - \frac{2}{nh}\sum_{i=1}^{n} K_h\left(x^1 - X_i^1\right)^2 U_i F'\left(\dot{M}_i\right) F''\left(\ddot{M}_i\right)^2 \left(\tilde{M}_i - M_i\right)^3 \\
&\quad + \frac{2}{nh}\sum_{i=1}^{n} K_h\left(x^1 - X_i^1\right)^2 U_i^2 F'\left(M_i\right) F''\left(\ddot{M}_i\right) \left(\tilde{M}_i - M_i\right) \\
&\quad + \frac{2}{nh}\sum_{i=1}^{n} K_h\left(x^1 - X_i^1\right)^2 F'\left(\dot{M}_i\right)^2 F'\left(M_i\right) F''\left(\ddot{M}_i\right) \left(\tilde{M}_i - M_i\right)^3 \\
&\quad - \frac{4}{nh}\sum_{i=1}^{n} K_h\left(x^1 - X_i^1\right)^2 U_i F'\left(\dot{M}_i\right) F'\left(M_i\right) F''\left(\ddot{M}_i\right) \left(\tilde{M}_i - M_i\right)^2 \\
&= T_1 + T_2 - 2T_3 + T_4 + T_5 - 2T_6 + 2T_7 + 2T_8 - 4T_9,
\end{aligned}$$

where $\dot{M}_i$ and $\ddot{M}_i$ are points on the line joining $\tilde{M}_i$ and M_i. For T_1, a similar argument to the proof of Lemma A.1 (i) yields $E\left[T_1\right] \to V_F$ and $Var\left(T_1\right) \to 0$. Thus, the Chebyshev's inequality implies $T_1 \xrightarrow{p} V_F$. From Horowitz and Mammen (2004, Theorem 1 (c)), we have

$$\max_{1\le i\le n}\left|\tilde{M}_i - M_i\right| = O_p\left(\kappa n^{-1/2} + \kappa^{-3/2}\right).$$

Thus, by applying the law of large numbers repeatedly, we can obtain $T_j \xrightarrow{p} 0$ for each $j = 2, \ldots, 9$. For example,

$$|T_2| \leq \left(\max_{1 \leq i \leq n} \left| \tilde{M}_i - M_i \right| \right)^2 \max_{1 \leq i \leq n} \left| F' \left(\dot{M}_i \right)^2 F' \left(M_i \right)^2 \right| \left| \frac{1}{nh} \sum_{i=1}^{n} K_h \left(x^1 - X_i^1 \right)^2 \right| \xrightarrow{p} 0,$$

where $\max_{1 \leq i \leq n} \left| F' \left(\dot{M}_i \right)^2 F' \left(M_i \right)^2 \right| = O_p(1)$ by Assumption 5 (i). Combining these results, the conclusion is obtained.

Proof of the second statement. Again, from the definition of $g_{F,i}$ and expansions around $\tilde{M}_i = M_i$,

$$\begin{aligned}
\frac{1}{\sqrt{nh}} \sum_{i=1}^{n} g_{F,i} &= \frac{1}{\sqrt{nh}} \sum_{i=1}^{n} K_h \left(x^1 - X_i^1 \right) \left\{ U_i - F' \left(\dot{M}_i \right) \left(\tilde{M}_i - M_i \right) \right\} \left\{ F' \left(M_i \right) + F'' \left(\ddot{M}_i \right) \left(\tilde{M}_i - M_i \right) \right\} \\
&= \frac{1}{\sqrt{nh}} \sum_{i=1}^{n} K_h \left(x^1 - X_i^1 \right) U_i F' \left(M_i \right) \\
&\quad + \frac{1}{\sqrt{nh}} \sum_{i=1}^{n} K_h \left(x^1 - X_i^1 \right) U_i F'' \left(\ddot{M}_i \right) \left(\tilde{M}_i - M_i \right) \\
&\quad - \frac{1}{\sqrt{nh}} \sum_{i=1}^{n} K_h \left(x^1 - X_i^1 \right) F' \left(\dot{M}_i \right) F' \left(M_i \right) \left(\tilde{M}_i - M_i \right) \\
&\quad - \frac{1}{\sqrt{nh}} \sum_{i=1}^{n} K_h \left(x^1 - X_i^1 \right) F' \left(\dot{M}_i \right) F'' \left(\ddot{M}_i \right) \left(\tilde{M}_i - M_i \right)^2 \\
&= L_1 + L_2 - L_3 - L_4,
\end{aligned}$$

where $\dot{M}_i$ and $\ddot{M}_i$ are points on the line joining $\tilde{M}_i$ and M_i. For L_1, Lyapunov's central limit theorem implies (note: $E[L_1] = 0$ by the law of iterated expectations)

$$L_1 \xrightarrow{d} N(0, V_F).$$

For L_2, by inserting the asymptotic linear form for $\left(\tilde{M}_i - M_i \right)$ given in Horowitz and Mammen (2004, Theorem 1 (d)), we can apply Horowitz and Mammen (2004, Lemma 8) to show that $L_2 \xrightarrow{p} 0$. Similar arguments yield $L_3 \xrightarrow{p} 0$ and $L_4 \xrightarrow{p} 0$. Therefore, the conclusion is obtained.

Proof of (ii). Since the proof is similar to Newey and Smith (2004, Lemmas A1 and A2), it is omitted.

A.4 Proof of Theorem 4.1

In this subsection, let f_1 be the density function of X^1, and

$$\begin{aligned}
g_{\tau,i} &= K_h \left(x^1 - X_i^1 \right) \left(\tau - I \left\{ \tilde{Y}_i \leq m_1(x^1) \right\} \right), \\
V_\tau &= \tau (1 - \tau) f_1 \left(x^1 \right) \int K(z)^2 \, dz.
\end{aligned}$$

From Lemma A.3 (ii), the first-order condition for $\hat{\lambda}_\tau$ satisfies

$$0 = \frac{1}{nh}\sum_{i=1}^{n}\frac{g_{\tau,i}}{1+\hat{\lambda}_\tau g_{\tau,i}} = \frac{1}{nh}\sum_{i=1}^{n} g_{\tau,i} - \hat{V}_{\tau,1}\hat{\lambda}_\tau,$$

w.p.a.1, where $\hat{V}_{\tau,1} = \frac{1}{nh}\sum_{i=1}^{n}\frac{g_{\tau,i}^2}{\left(1+\dot{\lambda}_\tau g_{\tau,i}\right)^2}$, the second equality follows from an expansion around $\hat{\lambda}_\tau = 0$, and $\dot{\lambda}_\tau$ is a point on the line joining $\hat{\lambda}_\tau$ and 0. Since

$$\left|\hat{V}_{\tau,1} - V_\tau\right| \leq \max_{1\leq i\leq n}\left|\frac{1}{1+\dot{\lambda}_\tau g_{\tau,i}}\right|^2 \left|\frac{1}{nh}\sum_{i=1}^{n} g_{\tau,i}^2 - V_\tau\right| \xrightarrow{p} 0,$$

(by Lemma A.3 (i) and (ii)) and $V_\tau > 0$ (by Assumptions 1 (iv) and 4 (i), and $\tau \in (0,1)$), $\hat{V}_{\tau,1}^{-1}$ exists w.p.a.1. Thus, we have

$$\hat{\lambda}_\tau = \hat{V}_{\tau,1}^{-1}\frac{1}{nh}\sum_{i=1}^{n} g_{\tau,i},$$

w.p.a.1. From Lemma A.3 (ii), $\hat{\lambda}_\tau$ satisfies $\ell_\tau\left(m_1(x^1)\right) = 2\sum_{i=1}^{n}\log\left(1+\hat{\lambda}_\tau g_{\tau,i}\right)$ w.p.a.1, and a second-order expansion of this equation around $\hat{\lambda}_\tau = 0$ yields

$$\ell_\tau\left(m_1(x^1)\right) = 2\hat{\lambda}_\tau\sum_{i=1}^{n} g_{\tau,i} - \hat{V}_{\tau,2}\hat{\lambda}_\tau^2 = \left[2\hat{V}_{\tau,1}^{-1} - \hat{V}_{\tau,2}\hat{V}_{\tau,1}^{-2}\right]\left(\frac{1}{\sqrt{nh}}\sum_{i=1}^{n} g_{\tau,i}\right)^2,$$

w.p.a.1, where $\hat{V}_{\tau,2} = \frac{1}{nh}\sum_{i=1}^{n}\frac{g_{\tau,i}^2}{\left(1+\ddot{\lambda}_\tau g_{\tau,i}\right)^2}$ and $\ddot{\lambda}_\tau$ is a point on the line joining $\hat{\lambda}_\tau$ and 0. Since $\left|\hat{V}_{\tau,2} - V_\tau\right| \xrightarrow{p} 0$ by the same argument to $\hat{V}_{\tau,1}$, we have $2\hat{V}_{\tau,1}^{-1} - \hat{V}_{\tau,2}\hat{V}_{\tau,1}^{-2} \xrightarrow{p} V_\tau^{-1}$. Therefore, Lemma A.3 (i) implies the conclusion.

Lemma A.3. *Under Assumptions 1-4 and 6 in Appendix A.1,*

(i) $\frac{1}{nh}\sum_{i=1}^{n} g_{\tau,i}^2 \xrightarrow{p} V_\tau$ *and* $\frac{1}{\sqrt{nh}}\sum_{i=1}^{n} g_{\tau,i} \xrightarrow{d} N\left(0, V_\tau\right)$*;*

(ii) *there exists* $\hat{\lambda}_\tau \in int\left(\Lambda_{\tau,n}\left(m_1(x^1)\right)\right)$ *satisfying*

$$\sum_{i=1}^{n}\log\left(1+\hat{\lambda}_\tau g_{\tau,i}\right) = \sup_{\lambda\in\Lambda_{\tau,n}(m_1(x^1))}\sum_{i=1}^{n}\log\left(1+\lambda g_{\tau,i}\right), \quad w.p.a.1,$$

$\left|\hat{\lambda}_\tau\right| = O_p\left((nh)^{-1/2}\right)$*, and* $\max_{1\leq i\leq n}\left|\hat{\lambda}_\tau g_{\tau,i}\right| \xrightarrow{p} 0.$

Proof of (i). Proof of the first statement. By the definition of $g_{\tau,i}$,

$$\begin{aligned}
\frac{1}{nh}\sum_{i=1}^{n} g_{\tau,i}^{2} &= \frac{1}{nh}\sum_{i=1}^{n} K_h\left(x^1 - X_i^1\right)^2\left(\tau - I\left\{\tilde{Y}_i \leq m_1(x^1)\right\}\right)^2 \\
&= \tau^2\frac{1}{nh}\sum_{i=1}^{n} K_h\left(x^1 - X_i^1\right)^2 + (1-2\tau)\frac{1}{nh}\sum_{i=1}^{n} K_h\left(x^1 - X_i^1\right)^2 I\left\{Y_i \leq m_1(x^1)\right\} \\
&\quad + (1-2\tau)\frac{1}{nh}\sum_{i=1}^{n} K_h\left(x^1 - X_i^1\right)^2\left(\Pr\left\{\tilde{Y}_i \leq m_1(x^1)\middle| X_i\right\} - \Pr\left\{Y_i \leq m_1(x^1)\middle| X_i\right\}\right) \\
&\quad + (1-2\tau)\frac{1}{nh}\sum_{i=1}^{n} K_h\left(x^1 - X_i^1\right)^2\begin{pmatrix} I\left\{\tilde{Y}_i \leq m_1(x^1)\right\} - \Pr\left\{\tilde{Y}_i \leq m_1(x^1)\middle| X_i\right\} \\ -I\left\{Y_i \leq m_1(x^1)\right\} + \Pr\left\{Y_i \leq m_1(x^1)\middle| X_i\right\}\end{pmatrix} \\
&= T_1 + T_2 + T_3 + T_3.
\end{aligned}$$

For T_1, a similar argument to the proof of Lemma A.1 (i) yields $E\left[T_1\right] \to \tau^2 f_1\left(x^1\right)\int K\left(z\right)^2 dz$ and $Var\left(T_1\right) \to 0$. Thus, the Chebyshev's inequality implies

$$T_1 \xrightarrow{p} \tau^2 f_1\left(x^1\right)\int K\left(z\right)^2 dz.$$

Similarly, for T_2, we obtain

$$T_2 \xrightarrow{p} \tau\left(1-2\tau\right) f_1\left(x^1\right)\int K\left(z\right)^2 dz.$$

By applying Horowitz and Lee (2005, Theorem 3 (a) and Lemma A.7), we can obtain $T_3 \xrightarrow{p} 0$. Also by applying Horowitz and Lee (2005, Theorem 3 (a) and Lemma A.5), we can obtain $T_4 \xrightarrow{p} 0$. Combining these results, the conclusion is obtained.

Proof of the second statement. Again, from the definition of $g_{\tau,i}$,

$$\begin{aligned}
\frac{1}{\sqrt{nh}}\sum_{i=1}^{n} g_{\tau,i} &= \frac{1}{\sqrt{nh}}\sum_{i=1}^{n} K_h\left(x^1 - X_i^1\right)\left(\tau - I\left\{\tilde{Y}_i \leq m_1(x^1)\right\}\right) \\
&= \frac{1}{\sqrt{nh}}\sum_{i=1}^{n} K_h\left(x^1 - X_i^1\right)\left(\tau - I\left\{Y_i \leq m_1(x^1)\right\}\right) \\
&\quad + \frac{1}{\sqrt{nh}}\sum_{i=1}^{n} K_h\left(x^1 - X_i^1\right)\left(\Pr\left\{Y_i \leq m_1(x^1)\middle| X_i\right\} - \Pr\left\{\tilde{Y}_i \leq m_1(x^1)\middle| X_i\right\}\right) \\
&\quad + \frac{1}{\sqrt{nh}}\sum_{i=1}^{n} K_h\left(x^1 - X_i^1\right)\begin{pmatrix} I\left\{Y_i \leq m_1(x^1)\right\} - \Pr\left\{Y_i \leq m_1(x^1)\middle| X_i\right\} \\ -I\left\{\tilde{Y}_i \leq m_1(x^1)\right\} + \Pr\left\{\tilde{Y}_i \leq m_1(x^1)\middle| X_i\right\}\end{pmatrix} \\
&= L_1 + L_2 + L_3.
\end{aligned}$$

For L_1, Lyapunov's central limit theorem implies (note: $E\left[L_1\right] = 0$ by the law of iterated expectations)

$$L_1 \xrightarrow{d} N\left(0, V_\tau\right).$$

By applying Horowitz and Lee (2005, Theorem 3 (a) and Lemma A.11), we can obtain $L_2 \xrightarrow{p} 0$. Also by applying Horowitz and Lee (2005, Theorem 3 (a) and Lemma A.13), we can obtain $L_3 \xrightarrow{p} 0$. Therefore, the conclusion is obtained.

Proof of (ii). Since the proof is similar to Newey and Smith (2004, Lemmas A1 and A2), it is omitted.

References

[1] Chen, S. X., Härdle, W. and M. Li (2003) An empirical likelihood goodness-of-fit test for time series, *Journal of the Royal Statistical Society*, B 65, 663-678.

[2] Chen, S. X. and Y. S. Qin (2000) Empirical likelihood confidence intervals for local linear smoothers, *Biometrika*, 87, 946-953.

[3] Chen, X. (2007) Large sample sieve estimation of semi-nonparametric models, In: Heckman, J. J. and E. E. Leamer (eds.) *Handbook of Econometrics*, vol. 6, part 2, 5469-5547, Elsevier, Amsterdam.

[4] De Gooijer, J. G. and D. Zerom (2003) On additive conditional quantiles with high-dimensional covariates, *Journal of the American Statistical Association*, 98, 135-146.

[5] Doksum, K. and J.-Y. Koo (2000) On spline estimators and prediction intervals in nonparametric regression, *Computational Statistics and Data Analysis*, 35, 76-82.

[6] Fan, J. and I. Gijbels (1996) *Local Polynomial Modelling and Its Applications*, Chapman & Hall, New York.

[7] Fan, J. and Jiang, J. (2005) Nonparametric inference for additive models, Journal of the American Statistical Association, 100, 890-907.

[8] Fan, J., Mammen, E. and W. Härdle (1998) Direct estimation of low dimensional components in additive models, *Annals of Statistics*, 26, 943-971.

[9] Fan, J., Zhang, C. and J. Zhang (2001) Generalized likelihood ratio statistics and Wilks phenomenon*, Annals of Statistics*, 29, 153-193.

[10] Buja, A., Hastie, T. J. and R. J. Tibshirani (1989) Linear smoothers and additive models, *Annals of Statistics*, 17, 453-555.

[11] Hastie, T. J. and R. J. Tibshirani (1990) *Generalized Additive Models*, Chapman & Hall, London.

[12] Hjort, N. L., McKeague, I. W. and I. van Keilegom (2009) Extending the scope of empirical likelihood, *Annals of Statistics*, 37, 1079-1111.

[13] Horowitz, J. L. (2001) Nonparametric estimation of a generalized additive model with an unknown link function, *Econometrica*, 69, 499-513.

[14] Horowitz, J., Klemelä, J. and E. Mammen (2006) Optimal estimation in additive regression models, *Bernoulli*, 12, 271-298.

[15] Horowitz, J. L. and S. Lee (2005) Nonparametric estimation of an additive quantile regression model, *Journal of the American Statistical Association*, 100, 1238–1249.

[16] Horowitz, J. L. and E. Mammen (2004) Nonparametric estimation of an additive model with a link function, *Annals of Statistics*, 32, 2412-2443.

[17] Kitamura, Y. (2007) Empirical likelihood methods in econometrics: theory and practice, in Blundell, R., Newey, W. K. and T. Persson (eds.), *Advances in Economics and Econometrics: Theory and Applications, Ninth World Congress*, Cambridge University Press.

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

[19] Koenker, R. and G. Bassett (1978) Regression quantiles, *Econometrica*, 46, 33-50.

[20] Linton, O. B. (1997) Efficient estimation of additive nonparametric regression models, *Biometrika*, 84, 469-473.

[21] Linton, O. B. and W. Härdle (1996) Estimating additive regression models with known links, *Biometrika*, 83, 529-540.

[22] Linton, O. B. and J. P. Nielsen (1995) A kernel method of estimating structured nonparametric regression based on marginal integration, *Biometrika*, 82, 93-100.

[23] Mammen, E., Linton, O. B. and J. P. Nielsen (1999) The existence and asymptotic properties of backfitting projection algorithm under weak conditions, *Annals of Statistics*, 27, 1443-1490.

[24] McCullagh, P. and J. Nelder (1989) *Generalized Linear Models*, 2nd ed., Chapman & Hall/CRC, Boca Raton.

[25] Newey, W. K. (1997) Convergence rates and asymptotic normality for series estimators, *Journal of Econometrics*, 79, 147-168.

[26] Newey, W. K. and R. J. Smith (2004) Higher order properties of GMM and generalized empirical likelihood estimators, *Econometrica*, 72, 219-255.

[27] Opsomer, J. D. (2000) Asymptotic properties of backfitting estimators, *Journal of Multivariate Analysis*, 73, 166-179.

[28] Opsomer, J. D. and D. Ruppert (1997) Fitting a bivariate additive model by local polynomial regression, *Annals of Statistics*, 25, 186-211.

[29] Otsu, T. (2008) Conditional empirical likelihood estimation and inference for quantile regression models, *Journal of Econometrics*, 142, 508-538.

[30] Otsu, T. and K.-L. Xu (2010) Empirical likelihood for regression discontinuity design, Working paper.

[31] Owen, A. B. (1988) Empirical likelihood ratio confidence intervals for a single functional, *Biometrika*, 75, 237-249.

[32] Owen, A. B. (2001) *Empirical Likelihood*, Chapman & Hall/CRC, New York.

[33] Qin, G. and Tsao, M. (2005) Empirical likelihood based inference for the derivative of the nonparametric regression function, *Bernoulli*, 11, 715-735.

[34] Qin, J. and J. Lawless (1994) Empirical likelihood and general estimating equations, *Annals of Statistics*, 22, 300-325.

[35] Stone, C. J. (1985) Additive regression and other nonparametric models, *Annals of Statistics*, 13, 689-705.

[36] Stone, C. J. (1986) The dimensionality reduction principle for generalized additive models, *Annals of Statistics*, 14, 590-606.

[37] Stone, C. J. (1994) The use of polynomial splines and their tensor products in multivariate function estimation (with discussion), *Annals of Statistics*, 22, 118-184.

[38] Wang, Q.-H. and B.-Y. Jing (2003) Empirical likelihood for partial linear models, *Annals of Institute of Statistical Mathematics*, 55, 585-595.