Ann Inst Stat Math (2010) 62:235–248
DOI 10.1007/s10463-008-0173-5

Boosting local quasi-likelihood estimators

Masao Ueki · Kaoru Fueda

Received: 13 March 2007 / Revised: 8 February 2008 / Published online: 25 April 2008

Abstract For likelihood-based regression contexts, including generalized linear models, this paper presents a boosting algorithm for local constant quasi-likelihood estimators. Its advantages are the following: (a) the one-boosted estimator reduces bias in local constant quasi-likelihood estimators without increasing the order of the variance, (b) the boosting algorithm requires only one-dimensional maximization at each boosting step and (c) the resulting estimators can be written explicitly and simply in some practical cases.

Keywords Bias reduction · L_2Boosting · Generalized linear models · Kernel regression · Local quasi-likelihood · Nadaraya–Watson estimator

1 Introduction

This paper deals with likelihood-based regression problems for which generalized linear models are typically used. However, the effectiveness of generalized linear models are limited because of their restricted flexibility. In the case, it is better to use some nonparametric approach such as kernel regression (Wand and Jones 1995; Fan and Gijbels 1996). Fan et al. (1995) extended the local constant and local polynomial regression estimators to quasi-likelihood methods, which is an extension of generalized linear models (see Sect. 2.1). Loader (1999) recommends the local quadratic fit that has the bias of $O(h^4)$ and the variance of $O\{(nh)^{-1}\}$, where h is the

M. Ueki (✉) · K. Fueda
Graduate School of Environmental Science, Okayama University,
Naka 3-1-1, Tsushima, Okayama 700-8530, Japan
e-mail: ueki@ems.okayama-u.ac.jp

K. Fueda
e-mail: fueda@ems.okayama-u.ac.jp

bandwidth, from a practical viewpoint. However, the local polynomial regression estimators require extensive computations because they rely on numerical maximization at each evaluated point. Fan (1999) overcomes that problem by introducing one-step local quasi-likelihood estimators, although some efforts at implementation are needed.

If one uses the local constant fit, such difficulties in both computation and implementation do not occur because it can be written explicitly and simply. However, the bias of the local constant fit is $O(h^2)$ which is often not negligible, while the variance is $O\{(nh)^{-1}\}$. We consequently take a course of not using local polynomials but applying a boosting algorithm to the local constant fit to reduce the order of the bias, where the boosting is a recently investigated statistical methodology (Schapire 1990; Freund 1995; Freund and Schapire 1996; Friedman 2001; Bühlmann and Yu 2003; Marzio and Taylor 2004a). Marzio and Taylor (2004b) proposed the boosting for Nadaraya–Watson estimator, which is the local constant fit in Gaussian model, where the algorithm they applied is the L_2Boosting of Friedman (2001) and Bühlmann and Yu (2003). The bias of the Nadaraya–Watson estimator is $O(h^2)$ and their one-boosted estimator reduces the bias to $O(h^4)$. This type of bias reduction is examined by many authors (Jones et al. 1995; Choi and Hall 1998; Marzio and Taylor 2004a).

The advantages of our algorithm are the following: (a) the one-boosted estimator reduces the bias of $O(h^2)$ to $O(h^4)$ without increasing the order of the variance, (b) our algorithm requires only one-dimensional maximization at each boosting step while the local polynomials need multi-dimensional maximization and (c) the resulting estimators can be written explicitly and simply in some practical cases. Our approach is also the simplest among the bias reduction techniques.

2 Boosting local constant quasi-likelihood estimators

2.1 Local constant quasi-likelihood estimators

This section describes local constant quasi-likelihood estimators. Let $(X_1, Y_1), \ldots, (X_n, Y_n)$ be a set of independent random pairs where, for each i, Y_i is a scalar response variable and X_i is an $\mathbb{R}^d$-valued vector of covariates having density f with support $\mathrm{supp}(f) \subseteq \mathbb{R}^d$. Let (X, Y) denote a generic member of the sample, and let $m(x) = E(Y|X = x)$. When the range of $m(x)$ is restricted on an interval I of $\mathbb{R}$ in likelihood based problems, with generalized linear models, such as Bernoulli, Poisson and gamma, the estimation is suitable for $\eta(x) = g\{m(x)\}$ instead of for $m(x)$ where g is a one-to-one function from I to $\mathbb{R}$, called link function.

The quasi-likelihood method is an extension of the generalized linear models. The former requires only the specification of a relationship between the mean and variance of Y; it is useful even if the likelihood function is not available. The former method maximizes the quasi-likelihood function $Q\{m(x), y\}$ instead of the log-likelihood function. This paper explains only the case in which the conditional variance is modeled as $\mathrm{var}(Y|X = x) = V\{m(x)\}$ for some known positive function V, and the corresponding quasi-likelihood function $Q(m, y)$ satisfies

$$\frac{\partial}{\partial m} Q(m, y) = \frac{y - m}{V(m)}. \tag{1}$$

The quasi-score (1) possesses properties that resemble those of the usual likelihood score function: one of the properties is that it satisfies the first two moment conditions of Bartlett's identities (Fan and Gijbels 1996). The likelihood score of one-parameter exponential family is a special case of (1) (Fan et al. 1995).

For simplicity, we deal with scalar covariates $X_1, \ldots, X_n$. The local constant quasi-likelihood estimator for $m(x)$ can be written explicitly as

$$\hat{m}_0(x; h) = g^{-1}\{\hat{\eta}_0(x; h)\} = \frac{\sum_{i=1}^n K_h(X_i - x)Y_i}{\sum_{i=1}^n K_h(X_i - x)}, \tag{2}$$

which is given by maximizing $\sum_{i=1}^n Q\{g^{-1}(\eta), Y_i\}K_h(X_i - x)$ with respect to η, where $K_h(z) = K(z/h)/h$, $K(z)$ is a symmetric unimodal probability density called kernel function, and $h > 0$ is a parameter called bandwidth, which controls the extent of smoothing. The estimator (2) is simple, but it performs poorly. In the next section, we strengthen (2) using boosting.

2.2 The boosting algorithm

In L_2boosting, a simple base estimator, called 'weak learner', is used iteratively in least-squares fitting with stage-wise updating of current residuals. In this section, before proposing the boosting local quasi-likelihood estimators, we describe the L_2boosting algorithm proposed by Marzio and Taylor (2004b), where the weak learner is the Nadaraya-Watson estimator, which corresponds to (2) when the link function g is the identity. The algorithm is given as follows.

Algorithm 1 *Step* 1 (*initialization*) *Let* $\hat{m}_0$ *be the Nadaraya-Watson estimator with a previously chosen* $h > 0$.

Step 2 (*iteration*) *Repeat for* $b = 0, \ldots, B$,

(i) *Compute* n *estimates* $\hat{m}_b(X_i)$, $i = 1, \ldots, n$.

(ii) *Update* $\hat{m}_{b+1}(x) = \hat{m}_b(x) + \hat{\delta}(x)$, *where* $\hat{\delta}(x)$ *is the Nadaraya-Watson estimator in which the response variables* Y_i *are replaced by the current residuals* $U_i = Y_i - \hat{m}_b(X_i)$, *i.e.,*

$$\hat{\delta}(x) = \frac{\sum_{i=1}^n K_h(X_i - x)U_i}{\sum_{i=1}^n K_h(X_i - x)}. \tag{3}$$

Least-squares fitting can be viewed as an optimization in the Gaussian regression model. This consideration enables us to generalize the L_2boosting to that in a quasi-likelihood framework. Here, we must take into account that additivity in step 2 (ii) does not necessarily hold in this framework. To achieve the generalization, we rewrite (3) as

$$\hat{\delta}(x) = \underset{\delta}{\operatorname{argmax}} \sum_{i=1}^n [Y_i - \{\hat{m}_b(X_i) + \delta\}]^2 K_h(X_i - x). \tag{4}$$

Based on the form of (4), we generalize Algorithm 1 for local constant quasi-likelihood estimators (2) as follows.

Algorithm 2 *Step* 1 (*initialization*) *Let* $\hat{\eta}_0$ *be* (2) *with a previously chosen* $h > 0$.
Step 2 (*iteration*) *Repeat for* $b = 0, \ldots, B$,

(i) *Compute n estimates* $\hat{\eta}_b(X_i)$, $i = 1, \ldots, n$.
(ii) *Update* $\hat{\eta}_{b+1}(x) = \hat{\eta}_b(x) + \hat{\delta}(x)$, *where*

$$\hat{\delta}(x) = \underset{\delta \in \mathbb{R}}{\operatorname{argmax}} \sum_{i=1}^{n} Q[g^{-1}\{\hat{\eta}_b(X_i) + \delta\}, Y_i] K_h(X_i - x). \tag{5}$$

We can obtain the estimator for $m(x)$ by $\hat{m}_{b+1}(x) = g^{-1}\{\hat{\eta}_{b+1}(x)\}$. Note that $\hat{\delta}(x)$ is added in η's space for range preservation, and (5) requires scalar maximization only, even for multiple covariates. Furthermore, some cases exist for which the resulting estimator can be written explicitly and simply as follows.

Example 1 (Gaussian model with identity link) This example corresponds to Algorithm 1. The quasi-likelihood function then coincides with the usual log-likelihood function of the Gaussian distribution with mean m and variance unity: $Q(m, y) = -\{(y-m)^2 + \log(2\pi)\}/2$; $V(m) = 1$. The link function g is the identity, $\eta = g(m) = m$. At the bth stage, $\hat{m}_{b+1}(x) = \hat{m}_b(x) + \hat{\delta}(x)$, where $\hat{\delta}(x)$ is given in (3).

Example 2 (Poisson model with log link) The link function g is log link, $\eta = g(m) = \log m$. The quasi-likelihood function then coincides with the usual log-likelihood function of the Poisson distribution with mean m: $Q(m, y) = -m + y \log m + \log y!$; $V(m) = m$. At the bth stage, $\hat{\eta}_{b+1}(x) = \hat{\eta}_b(x) + \hat{\delta}(x)$, where

$$\exp\{\hat{\delta}(x)\} = \frac{\sum_{i=1}^{n} K_h(X_i - x) Y_i}{\sum_{i=1}^{n} K_h(X_i - x) \exp\{\hat{\eta}_b(X_i)\}}.$$

Example 3 (gamma model with log link) The link function g is log link, $\eta = g(m) = \log m$. The quasi-likelihood function then coincides with the usual log-likelihood function of the gamma density with mean m and shape parameter α: $Q(m, y) = -\alpha y/m - \alpha \log m + (\alpha - 1) \log y + \alpha \log \alpha - \log \Gamma(\alpha)$, where $\Gamma(\cdot)$ is the gamma function; $V(m) = m^2/\alpha$. At the bth stage, $\hat{\eta}_{b+1}(x) = \hat{\eta}_b(x) + \hat{\delta}(x)$, where

$$\exp\{\hat{\delta}(x)\} = \frac{\sum_{i=1}^{n} K_h(X_i - x) \exp\{-\hat{\eta}_b(X_i)\} Y_i}{\sum_{i=1}^{n} K_h(X_i - x)}.$$

2.3 Emphasizing the updating term

Primarily, boosting can be regarded as a sequential greedy optimization of additive models, which is typical in L_2boosting. The updating term in each step of boosting can be regarded as an estimation using iteratively reweighted data. From this viewpoint, we specifically examine the updating term defined in (5).

Second-order Taylor approximation in (5) yields that

$$\begin{aligned}\ell_n(\delta) =& \sum_{i=1}^{n} K\left(\frac{X_i - x}{h}\right) Q[g^{-1}\{\hat{\eta}_b(X_i) + \delta, Y_i\}] \\ \approx& \sum_{i=1}^{n} K\left(\frac{X_i - x}{h}\right)\left(Q[g^{-1}\{\hat{\eta}_b(X_i)\}, Y_i] + \delta q_1\{\hat{\eta}_b(X_i), Y_i\} + \frac{1}{2}\delta^2 q_2\{\hat{\eta}_b(X_i), Y_i\}\right),\end{aligned}$$

where q_i are defined in the Appendix. Therefore, the updating term is approximated as the following.

$$\begin{aligned}\hat{\delta}(x) \approx& \frac{\sum_{i=1}^{n} K\left(\frac{X_i - x}{h}\right) q_1\{\hat{\eta}_b(X_i), Y_i\}}{-\sum_{i=1}^{n} K\left(\frac{X_i - x}{h}\right) q_2\{\hat{\eta}_b(X_i), Y_i\}} \\ =& \frac{\sum_{i=1}^{n} K_h(X_i - x)[g'\{\hat{m}_b(X_i)\}V\{\hat{m}_b(X_i)\}]^{-1}\{Y_i - \hat{m}_b(X_i)\}}{-\sum_{i=1}^{n} K_h(X_i - x) q_2\{\hat{\eta}_b(X_i), Y_i\}}. \qquad (6)\end{aligned}$$

Using (6), the updating term $\hat{\delta}(x)$ can be interpreted approximately as the reweighted version of the kernel regressor (5), where the response variables Y_i are replaced by the current residuals as in (3) in Algorithm 1. This consideration describes a transparent relationship between the proposed algorithm and L_2boosting.

3 Bias reduction property

In the following theorem, we state the bias reduction property, i.e., one-boosted estimators $\hat{\eta}_1(x; h)$ reduce the bias of $O(h^2)$ in local constant quasi-likelihood estimators, which is often not negligible, to $O(h^4)$.

Theorem 1 *Suppose that the conditions presented in the Appendix hold. If $h \to 0$ and $nh \to \infty$ as $n \to \infty$, the estimator after one-boosting iteration, $\hat{\eta}_1(x; h)$, has bias of $O(h^4)$ and variance of*

$$\mathrm{var}\{\hat{\eta}_1(x; h)\} = \mathrm{var}(Y|X = x)\frac{g'\{m(x)\}^2}{nhf(x)} \int T_K^2(z)dz + o\{(nh)^{-1}\},$$

where $T_K(z) = 2K(z) - K * K(z)$ is the fourth order kernel in Jones et al. (1995, Theorem 1). In addition, $\hat{m}_1(x; h) = g^{-1}\{\hat{\eta}_1(x; h)\}$ has the bias of $O(h^4)$ and the variance of $\mathrm{var}\{\hat{m}_1(x; h)\} = \mathrm{var}\{\hat{\eta}_1(x; h)\}/g'\{m(x)\}^2 + o\{(nh)^{-1}\}$.

The proof is given in the Appendix. According to Fan et al. (1995), the bias and variance of the local constant quasi-likelihood estimators, i.e., the non-boosted estimators, are $O(h^2)$ and $O\{(nh)^{-1}\}$, respectively, which in turn implies that bias reduction is achieved without increasing the order of the variance.

4 Numerical illustrations

This section provides some numerical illustrations in Poisson (example 2) and exponential (Example 3 for $\alpha = 1$) models. We use the Epanechnikov kernel $K(u) = \frac{3}{4}(1-u^2)1_{\{-1<u<1\}}$, where $1_{\{\cdot\}}$ is the indicator function. The examined conditional means are

$$m_{p1}(x) = \exp\{\cos(2\pi x)\}, \quad m_{p2}(x) = \arcsin x + 2, \quad \text{for Poisson,}$$
$$m_{e1}(x) = 8\exp(-x^2), \quad m_{e2}(x) = 3(x+1)^{1/2} + 4, \quad \text{for exponential,}$$

and the design density $f(x)$ is the uniform density on $[-1, 1]$. To measure the performance of resulting estimator $\hat{m}(x)$, we use the square root of average square errors, $\text{RASE} = \left[\frac{1}{300}\sum_{j=1}^{300}\{m(x_j) - \hat{m}(x_j)\}^2\right]^{1/2}$, for $x_j = -1 + 2(j-1)/299$, $j = 1, \ldots, 300$, at which the function $m(x)$ is estimated. The sample size n is 100 throughout. To show how the proposed algorithm works, we demonstrate the behaviors for one random sample in Fig. 1 (Poisson) and Fig. 2 (exponential) where $\hat{m}_b(x)$ are plotted for $b = 0$ (dash), 1 (dot), 2 (dot dash) and 3 (long dash), together with the true curve (solid). The bandwidth h used in the left and right panels are optimal, respectively, for $b = 0$ and $b = 3$, which are founded numerically with respect to the RASE. It seems that $\hat{m}_3(x)$ in the right panels fit more appropriately to the true curves than the $\hat{m}_0(x)$ in the left panels: the optimal boosted estimators are better than the optimal non-boosted ones.

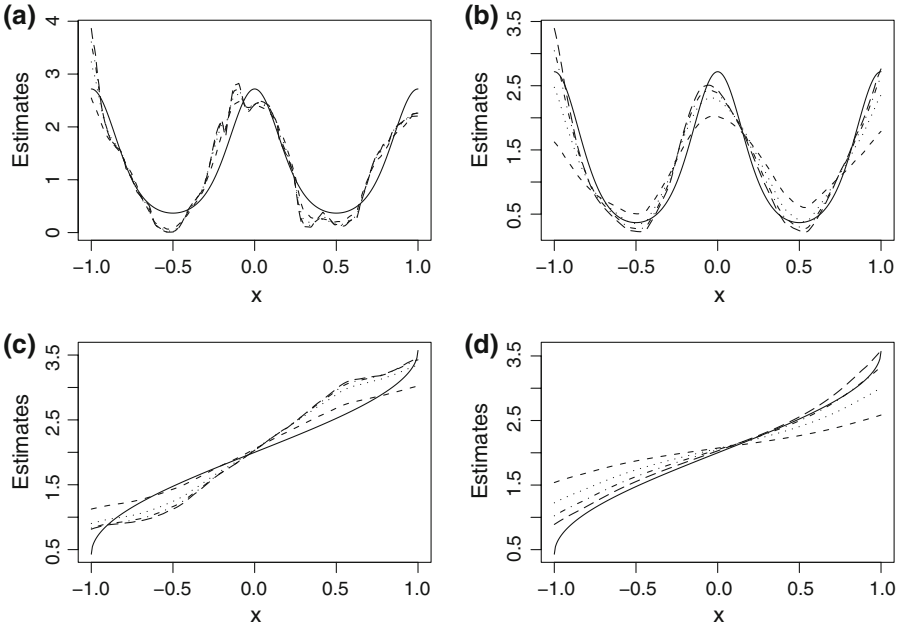

Fig. 1 Estimates in a Poisson case: for $m_{p1}(x)$, **a** $h = 0.17$, **b** $h = 0.42$; for $m_{p2}(x)$, **c** $h = 0.96$, **d** $h = 1.67$

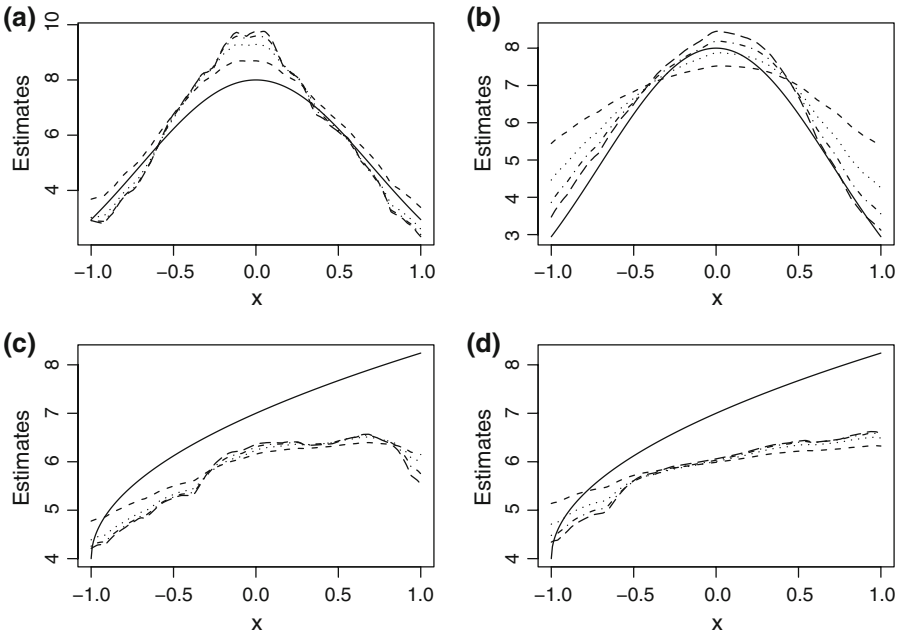

Fig. 2 Estimates in an exponential case: for $m_{e1}(x)$, **a** $h = 0.54$, **b** $h = 0.99$; for $m_{e2}(x)$, (c) $h = 0.97$, **d** $h = 1.26$

We examine the boosting for various h, and repeat the procedures 500 times to illustrate the efficiency. Figure 3 shows the average RASEs against h, where the plotted numbers 0–3 indicate the corresponding boosting iterations (the 0 corresponds to non-boosted estimator, i.e., local constant quasi-likelihood estimator). All figures suggest that boosting works well for appropriate h because each minimum RASE of the boosted estimate is smaller than that of non-boosted estimate. Note that h which minimizes the RASE tends to increase as the number of boosting iterations grows. This phenomenon is identical to that observed in Marzio and Taylor (2004a) for boosting kernel density estimators. Therefore, we recommend to take somewhat larger h than the optimal one for non-boosted estimators as the strategy to select h.

Next, we verify the implication in Theorem 1 related to the mean squared error (MSE). Table 1 compares theoretical MSE expressions given in Theorem 1 and simulated true MSEs in 1000 experiments, where both the non-boosted and one-boosted estimators, $\hat{m}_0(x)$ and $\hat{m}_1(x)$ are evaluated at three points $x = -0.3, 0, 0.6$. The results show that the asymptotic MSE expressions given in Theorem 1 approximate the true MSEs well.

In practice, the bandwidth h must be estimated from the data. One way of choosing h is to use likelihood based cross-validation. The method is useful when the form of $Q(m, y)$ is known, as in generalized linear models. The bandwidth $\hat{h}$ selected by the cross-validation is the h maximizing

$$\sum_{i=1}^{n} Q\{\hat{m}_{-i}(X_i), Y_i\},$$

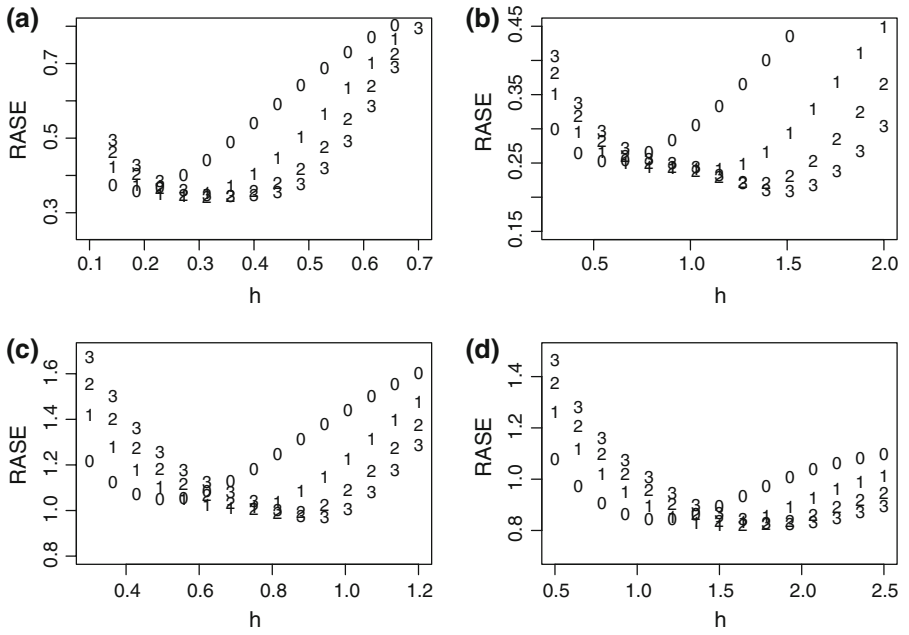

Fig. 3 Average RASE plots: **a** for $m_{p1}(x)$, **b** for $m_{p2}(x)$, **c** for $m_{e1}(x)$ and **d** for $m_{e2}(x)$. The plotted numbers 0–3 indicate the corresponding boosting iterations

Table 1 Theoretical MSE expressions and simulated MSEs for non-boosted and one-boosted estimators, $\hat{m}_0(x)$ and $\hat{m}_1(x)$

T/S	B	$m_{p1}(h = 0.2)$	$m_{p2}(h = 1.2)$	$m_{e1}(h = 0.8)$	$m_{e2}(h = 1.7)$
T	0	0.052, 0.628, 0.018	0.031, 0.040, 0.098	1.39, 2.01, 0.51	0.44, 0.39, 0.44
S	0	0.088, 0.294, 0.040	0.030, 0.025, 0.097	1.19, 1.65, 0.77	0.49, 0.49, 0.98
T	1	0.062, 0.231, 0.038	0.024, 0.028, 0.037	1.14, 1.36, 0.66	0.42, 0.49, 0.61
S	1	0.083, 0.266, 0.039	0.024, 0.025, 0.057	1.18, 1.32, 0.76	0.44, 0.49, 0.90

In the table, each MSE evaluated at $x = -0.3, 0, 0.6$ is described in the order corresponding to that of x. 'T' and 'S' denote Theoretical and Simulated values, respectively. B means the boosting iteration number

where $\hat{m}_{-i}(\cdot)$ corresponds to the version of $\hat{m}(\cdot)$ that is constructed by eliminating ith data (X_i, Y_i). Table 2 shows the average RASEs in 500 simulation experiments for respective boosting iteration numbers 0–3, with bandwidths selected using cross-validation. The bandwidths selected here are chosen among finite candidates, which consist of 50 equi-spaced points on the intervals given in the second line of Table 2. These intervals are determined empirically according to the variability of $\hat{h}$. By the results in Table 2, we ascertained that the boosted estimation, at least once, works better than non-boosted estimation, even if the bandwidths are estimated using cross-validation. Consequently, it is worthwhile to apply the boosting algorithm in practical situations.

Table 2 Average RASE for each boosting number, with bandwidth selected by cross-validation

B	m_{p1} [0.1,1.5]	m_{p2} [0.2,2]	m_{e1} [0.13,2.3]	m_{e2} [0.1,4.5]
0	0.376	0.282	1.274	1.090
1	0.372	0.278	1.263	1.062
2	0.369	0.273	1.263	1.076
3	0.370	0.269	1.255	1.108

B means the boosting iteration number. The candidate bandwidths consist of 50 equi-spaced points on the intervals given in the second line of the table

5 Concluding remarks

We propose a boosting algorithm for local constant quasi-likelihood estimators that provides bias reduction. The method is valid in both computation and implementation. There are still some issues. The first is the selection of h. A reasonable solution in generalized linear models is to use likelihood-based cross-validation. In some cases in which the resulting estimators are given explicitly, the required computations for cross-validation are few. However, in other cases in which numerical maximizations are required, including logistic regression, the required computations could be expensive. For this reason, better selection criteria are needed. The second is how to stop the boosting iteration. In our examinations, the two-boosted and three-boosted estimators work better than the one-boosted estimators. However, as Bühlmann and Yu (2003) pointed out, many boosting iterations cause overfitting. To avoid this, we have to stop the iteration based on a stopping rule such as cross-validation. The third is to analyze the two-boosted and more-boosted estimators because we have only justified the one-boosted estimators in this paper.

Acknowledgments The authors would like to thank the referees for helpful suggestions that improve the paper considerably.

Appendix: Proof of Theorem 1

Preliminary Let $q_i(\eta, y) = (\partial^i/\partial\eta^i)Q\{g^{-1}(\eta), y\}$ for $i = 1, 2$. Since Q satisfies (1), q_i is linear in y for fixed x, $q_1\{\eta(x), m(x)\} = 0$ and $q_2\{\eta(x), m(x)\} = -\rho(x)$, where $\rho(x) = [g'\{m(x)\}^2 V\{m(x)\}]^{-1}$. Also let $\sigma^2(x) = \mathrm{var}(Y|X = x)$.

We present the conditions: (i) The function $q_2(\eta, y) < 0$ for $\eta \in \mathbb{R}$ and y in the range of the response variable; (ii) The functions $f^{(4)}$, $\eta^{(4)}$, σ^2, V'' and $g^{(4)}$ are continuous; (iii) For each $x \in \mathrm{supp}(f)$, $\rho(x)$, $\sigma^2(x)$ and $g'\{m(x)\}$ are nonzero; (iv) The kernel K is a symmetric probability density with support $[-1, 1]$; (v) x is an interior point of $\mathrm{supp}(f)$. Furthermore, we assume that $h \propto n^{-1/9}$, which is the optimal rate that minimizes the asymptotic MSE of order $O\{h^8 + (nh)^{-1}\}$. See the argument below Theorem 1 of Jones et al. (1995). We also write $(f\rho)(x) = f(x)\rho(x)$ and $(mf)(x) = m(x)f(x)$.

Let $\delta^* = a_n^{-1}\delta$, $\hat{\eta}_i(x) = \hat{\eta}_i(x;h)$ for $i = 0, 1$, $\hat{m}_0(x) = \hat{m}_0(x;h)$ and

$$\ell_n(\delta^*) = \sum_{i=1}^{n} K\left(\frac{X_i - x}{h}\right)\left(Q[g^{-1}\{\hat{\eta}_0(X_i) + a_n\delta^*\}, Y_i] - Q[g^{-1}\{\hat{\eta}_0(X_i)\}, Y_i]\right),$$

where $a_n = (nh)^{-1/2}$. Condition (i) implies that ℓ_n is concave in δ^*. Let $\hat{\delta}^*$ be the maximizer of $\ell_n(\delta^*)$, then

$$\hat{\delta}^* = \frac{1}{(f\rho)(x)} W_n + o_p(1) \quad \text{where} \quad W_n = a_n \sum_{i=1}^{n} K\left(\frac{X_i - x}{h}\right) q_1\{\hat{\eta}_0(X_i), Y_i\}. \tag{7}$$

The derivation of (7) is as follows. Using Taylor expansion,

$$\ell_n(\delta^*) = W_n\delta^* + \frac{1}{2}A_n\delta^{*2} + \frac{a_n^3}{6}\sum_{i=1}^{n} K\left(\frac{X_i - x}{h}\right) q_3(\eta_i, Y_i)\delta^{*3}, \tag{8}$$

where η_i is between $\hat{\eta}_0(X_i)$ and $\hat{\eta}_0(X_i) + a_n\delta^*$, and $A_n = a_n^2 \sum_{i=1}^{n} K\left(\frac{X_i - x}{h}\right)$ $q_2\{\hat{\eta}_0(X_i), Y_i\}$. By $\hat{\eta}_0(x) = \eta(x) + o_p(1)$,

$$E(A_n) = h^{-1}E[K\left(\frac{X_1 - x}{h}\right) q_2\{\eta(X_1), m(X_1)\}] + o(1) = -(f\rho)(x) + o(1)$$

and $\mathrm{var}(A_n) = O(a_n^2)$; using $A_n = E(A_n) + O_p\{\mathrm{var}(A_n)^{1/2}\}$, we have $A_n = -(f\rho)(x) + o_p(1)$. A similar argument in Fan and Gijbels (1996, p. 212) shows that the last term in (8) is bounded by $O_p(a_n)$. Therefore, $\ell_n(\delta^*) = W_n\delta^* - \frac{1}{2}(f\rho)(x)\delta^{*2} + o_p(1)$. Using the quadratic approximation lemma (Fan and Gijbels 1996, p. 210), we obtain (7).

Bias First, we derive the bias. Let $\mu_2 = \int z^2 K(z)dz$. Using $\frac{1}{n}\sum_{i=1}^{n} K_h(X_i - x) = f(x) + \frac{1}{2}h^2\mu_2 f''(x) + O_p(h^4)$ by conditions (ii), (iv) and $(nh)^{-1/2} \propto n^{-4/9} \propto h^4$, it follows from (2) that

$$g^{-1}\{\hat{\eta}_0(x)\} = \hat{m}_0(x) = \frac{1}{nf(x)}\sum_{j=1}^{n} K_h(X_j - x)Y_j\left\{1 - \frac{1}{2}h^2\mu_2\frac{f''(x)}{f(x)}\right\} + O_p(h^4). \tag{9}$$

Using $q_1\{\eta(x), y\} = g'\{m(x)\}\rho(x)[y - g^{-1}\{\eta(x)\}]$, (7) is rewritten as

$$\hat{\delta}^* = \frac{a_n h}{(f\rho)(x)} \sum_{i=1}^{n} K_h(X_i - x)\rho(X_i)g'\{m(X_i)\} \times \left[Y_i - \frac{1}{nf(X_i)} \sum_{j \neq i} K_h(X_j - X_i)Y_j \left\{ 1 - \frac{1}{2}h^2\mu_2 \frac{f''(X_i)}{f(X_i)} \right\} \right] + O_p(a_n^{-1}h^4). \tag{10}$$

In addition, using conditions (ii), (iv) and $\int K_h(v-u)(mf)(v)dv = (mf)(u) + \frac{1}{2}h^2\mu_2(mf)''(u) + O(h^4)$,

$$\begin{aligned} E(\hat{\delta}^*) &= \frac{a_n nh}{(f\rho)(x)} \int (f\rho)(u)K_h(u-x)g'\{m(u)\} \\ &\quad \times \left[m(u) - \frac{1}{f(u)} \int K_h(v-u)(mf)(v)\mathrm{d}v \left\{ 1 - \frac{1}{2}h^2\mu_2 \frac{f''(u)}{f(u)} \right\} \right] \\ &\quad \times \mathrm{d}u + O(a_n^{-1}h^4) \\ &= -\frac{a_n^{-1}}{(f\rho)(x)} \int (f\rho)(u)K_h(u-x)g'\{m(u)\}h^2\mu_2 \\ &\quad \times \left\{ \frac{(mf)''(u)}{2f(u)} - \frac{(mf'')(u)}{2f(u)} \right\} \mathrm{d}u + O(a_n^{-1}h^4) \\ &= -a_n^{-1} \frac{g'\{m(x)\}}{2f(x)} h^2\mu_2 \left\{ (mf)''(x) - (mf'')(x) \right\} + O(a_n^{-1}h^4). \end{aligned} \tag{11}$$

According to Fan et al. (1995), the bias of $\hat{\eta}_0(x)$ is given as

$$E\{\hat{\eta}_0(x)\} - \eta(x) = \frac{g'\{m(x)\}}{2f(x)} h^2\mu_2\{(mf)''(x) - (mf'')(x)\} + O(h^4). \tag{12}$$

Combining (11) and (12), we can show that the bias of $\hat{\eta}_1(x) = \hat{\eta}_0(x) + \hat{\delta}(x)$ is $O(h^4)$.

Variance Secondly, we derive the variance. Define $\hat{\eta}_i^*(x) = a_n^{-1}[\hat{\eta}_i(x) - E\{\hat{\eta}_i(x)|\mathbb{X}\}]$ for $i = 0, 1$, where $E\{\cdot|\mathbb{X}\}$ is the conditional expectation under given $X_1, \ldots, X_n$. Then, it holds that $\mathrm{var}\{\hat{\eta}_1(x)\} = a_n^2 E\{\hat{\eta}_1^*(x)^2\} + E\left([E\{\hat{\eta}_1(x)|\mathbb{X}\} - \eta(x)]^2\right) - [E\{\hat{\eta}_1(x)\} - \eta(x)]^2$, where the third term, the squared bias, is $\{O(h^4)\}^2$. To calculate the second term, we first note, using the Taylor expansion for (9), that

$$\hat{\eta}_0(x) - \eta(x) = \frac{g'\{m(x)\}}{nf(x)} \sum_{j=1}^{n} \left[K_h(X_j - x)Y_j \left\{ 1 - \frac{1}{2}h^2\mu_2 \frac{f''(x)}{f(x)} \right\} - (mf)(x) \right] + O_p(h^4). \tag{13}$$

From (10) and (13),

$$E\{\hat{\eta}_1(x)|\mathbb{X}\} - \eta(x) = E\{\hat{\eta}_0(x) - \eta(x)|\mathbb{X}\} + E\{\hat{\delta}(x)|\mathbb{X}\} = D + O_p(h^4), \quad (14)$$

where $D = \frac{1}{n}\sum_{i=1}^n R_i + \frac{1}{n^2}\sum_{i\neq j}^n S_{ij}$, $R_i = R(X_i)$, $S_{ij} = S(X_i, X_j)$,

$$\begin{aligned} R(X_i) &= \frac{g'\{m(x)\}}{f(x)}\left[K_h(X_i - x)m(X_i)\left\{1 - \frac{1}{2}h^2\mu_2\frac{f''(x)}{f(x)}\right\} - (mf)(x)\right] \\ &\quad + \frac{1}{(f\rho)(x)}K_h(X_i - x)\rho(X_i)g'\{m(X_i)\}m(X_i), \\ S(X_i, X_j) &= -\frac{1}{(f\rho)(x)}K_h(X_i - x)\frac{\rho(X_i)g'\{m(X_i)\}}{f(X_i)}K_h(X_j - X_i)m(X_j) \\ &\quad \times\left\{1 - \frac{1}{2}h^2\mu_2\frac{f''(X_i)}{f(X_i)}\right\}. \end{aligned}$$

Note that $E(D)$ equals the bias of $\hat{\eta}_1(x)$ with error $O(h^4)$, i.e., $E(D) = O(h^4)$. Observing that $E_1(R_1)$ and $E_{12}(S_{12})$ are of order $O(1)$,

$$\begin{aligned} E(D^2) &= E\left(\frac{1}{n^2}\sum_{i,j}R_iR_j + \frac{2}{n^3}\sum_i\sum_{j\neq k}R_iS_{jk} + \frac{1}{n^4}\sum_{i\neq j}\sum_{k\neq l}S_{ij}S_{kl}\right) \\ &= \{E_1(R_1)\}^2 + 2E_1(R_1)E_{12}(S_{12}) + \{E_{12}(S_{12})\}^2 + O(n^{-1}) \\ &= \{E_{12}(R_1 + S_{12})\}^2 + O(n^{-1}) = \{E(D)\}^2 + O(n^{-1}) = O(h^8), \end{aligned}$$

where E_1 and E_{12} represent expectations with respect to X_1 and (X_1, X_2), respectively. Therefore, the second term is also of order $O(h^8)$.

It is, after all, sufficient to calculate $E\{\hat{\eta}_1^*(x)^2\}$. By (13) and $(na_n)^{-1} = a_nh$,

$$\hat{\eta}_0^*(x) = a_nh\frac{g'\{m(x)\}}{f(x)}\sum_{j=1}^n K_h(X_j - x)\tilde{Y}_j + O_p(h^2), \quad (15)$$

in which $\tilde{Y}_i = Y_i - m(X_i)$. On the other hand, defining $G_{r,n} = a_nh\sum_{i=1}^n K_h(X_i - x)\left\{\tilde{Y}_i - \frac{1}{nf(X_i)}\sum_{j\neq i}K_h(X_j - X_i)\tilde{Y}_j\right\}(X_i - x)^r$ for $r = 0, 1$ and $\xi(x) = \rho(x)g'\{m(x)\}$, it follows from Taylor expanding $\xi(X_i)$ around x in (10), with condition (ii), that

$$\begin{aligned} \hat{\delta}^* - E(\hat{\delta}^*|\mathbb{X}) &= \frac{1}{(f\rho)(x)}\{G_{0,n}\xi(x) + G_{1,n}\xi'(x)\} + O_p(h^2) \\ &= \frac{g'\{m(x)\}}{f(x)}G_{0,n} + o_p(1). \end{aligned} \quad (16)$$

The second equality follows from $G_{1,n} = o_p(1)$, which we show in what follows. Observing that $E(\tilde{Y}_i\tilde{Y}_j) = 0$ if $i \neq j$ and $= \int \sigma^2(w)f(w)dw$ otherwise, we have

$$\begin{aligned}
&\frac{1}{(a_n h)^2}E(G_{r,n}^2)\\
&= E\Bigg[\sum_{i,k=1}^{n} K_h(X_i - x)K_h(X_k - x)\Bigg\{\tilde{Y}_i\tilde{Y}_k - \frac{2}{nf(X_i)}\sum_{j\neq i} K_h(X_j - X_i)\tilde{Y}_j\tilde{Y}_k\\
&\quad + \frac{1}{n^2 f(X_i)f(X_k)}\sum_{j\neq i, l\neq k} K_h(X_j - X_i)K_h(X_l - X_k)\tilde{Y}_j\tilde{Y}_l\Bigg\}(X_i - x)^r(X_k - x)^r\Bigg]\\
&= I_1 - 2I_2 + I_3 + o(I_1 - 2I_2 + I_3),
\end{aligned}$$

where

$$\begin{aligned}
I_1 &= n\int K_h^2(w - x)(f\sigma^2)(w)(w - x)^{2r}\mathrm{d}w,\\
I_2 &= n\int K_h(w - x)\int K_h(u - x)K_h(u - w)(f\sigma^2)(u)(u - x)^r du(w - x)^r\mathrm{d}w,\\
I_3 &= n\int K_h(u - x)\int K_h(v - x)\int K_h(w - u)K_h(w - v)(f\sigma^2)(w)\\
&\qquad \times \mathrm{d}w(u - x)^r(v - x)^r\mathrm{d}u\mathrm{d}v.
\end{aligned}$$

Then, $I_1 = nh^{2r-2}\int z^{2r}K^2(z)(f\sigma^2)(x + hz)h\mathrm{d}z = O(a_n^{-2}h^{2r-2})$ and

$$\begin{aligned}
I_2 &= nh^{2r-3}\int K(z)\int K\left(\frac{u - x}{h}\right)K\left(\frac{u - x}{h} - z\right)(f\sigma^2)(u)\left(\frac{u - x}{h}\right)^r \mathrm{d}u z^r h\mathrm{d}z\\
&= nhh^{2r-2}\int K(z)\int K(s)K(s - z)(f\sigma^2)(x + sh)s^r\mathrm{d}s z^r\mathrm{d}z = O(a_n^{-2}h^{2r-2}).
\end{aligned}$$

Similarly,

$$\begin{aligned}
I_3 &= nh^{2r-4}\int K(s)\int K(t)\int K\left(\frac{w - x}{h} - s\right)K\left(\frac{w - x}{h} - t\right)(f\sigma^2)(w)\mathrm{d}w\, sh\mathrm{d}s\, th\mathrm{d}t\\
&= nhh^{2r-2}\int K(s)\int K(t)\int K(z - s)K(z - t)(f\sigma^2)(x + hz)\mathrm{d}z s\mathrm{d}s t\mathrm{d}t\\
&= O(a_n^{-2}h^{2r-2}).
\end{aligned}$$

Thus, we deduce that $E(G_{r,n}^2) = O(h^{2r})$. Noting that $E(G_{r,n}) = 0$, $\mathrm{var}(G_{r,n}) = E(G_{r,n}^2)$. This implies that $G_{r,n} = O_p(\sqrt{h^{2r}}) = O_p(h^r)$, in particular, $G_{1,n} = o_p(1)$, thereby yielding (16).

Combining (15) and (16), we can write

$$\hat{\eta}_1^*(x) = \frac{g'\{m(x)\}}{f(x)} Z_n + o_p(1), \tag{17}$$

where $Z_n = a_n h \sum_{i=1}^n K_h(X_i - x) \left\{ 2\tilde{Y}_i - \frac{1}{nf(X_i)} \sum_{j \neq i} K_h(X_j - X_i)\tilde{Y}_j \right\}$; consequently, $E\{\hat{\eta}_1^*(x)^2\} = \left[\frac{g'\{m(x)\}}{f(x)} \right]^2 E(Z_n^2) + o(1)$. The same arguments as in deriving (16) apply to the calculation of $E(Z_n^2)$, which derives the variance.

Bias and variance of $\hat{m}_1(x)$ The assertion regarding the estimator for $m(x)$ is straightforwardly obtained by noting $\hat{m}_1(x) = g^{-1}\{\hat{\eta}_1(x)\}$ and using the same process as that of the proof of Theorem 2 in Fan et al. (1995).

References

Bühlmann, P., Yu, B. (2003). Boosting with the L_2 loss: regression and classification. *Journal of the American Statistical Association*, *98*, 324–339.

Choi, E., Hall, P. (1998). On bias reduction in local linear smoothing. *Biometrika*, *85*, 333–345.

Fan, J. (1999). One-step local quasi-likelihood estimation. *Journal of the Royal Statistical Society, Ser. B*, *61*, 927–943.

Fan, J., Gijbels, I. (1996). *Local polynomial modelling and its applications.* London: Chapman and Hall.

Fan, J., Heckman, N. E., Wand, M. P. (1995). Local polynomial kernel regression for generalized linear models and quasi-likelihood functions. *Journal of the American Statistical Association*, *90*, 141–150.

Freund, Y. (1995). Boosting a weak learning algorithm by majority. *Information and Computation*, *121*, 256–285.

Freund, Y., Schapire, R. E. (1996). Experiments with a new boosting algorithm. In: Saitta, L. (Ed.) *Machine Learning: Proceedings of the Thirteenth International Conference,* (pp. 144–156). San Francisco: Morgan Kauffman.

Friedman, J. (2001). Greedy function approximation: a gradient boosting machine. *The Annals of Statistics*, *29*, 1189–1232.

Jones, M. C., Linton, O. and Nielsen, J. (1995). A simple bias reduction method for density estimation. *Biometrika*, *82*, 327–338.

Loader, C. R. (1999). *Local regression and likelihood.* New York: Springer.

Marzio, M. D., Taylor, C. C. (2004a). Boosting kernel density estimates: A bias reduction technique? *Biometrika*, *91*, 226–233.

Marzio, M. D., Taylor, C. C. (2004b). Multistep kernel regression smoothing by boosting. www.amsta.leeds.ac.uk/~charles/boostreg.pdf, unpublished manuscript.

Schapire, R. E. (1990). The strength of weak learnability. *Machine Learning*, *5*, 313–321.

Wand, M. P., Jones, M. C. (1995). *Kernel smoothing.* London: Chapman and Hall.