The Annals of Statistics
1999, Vol. 27, No. 1, 36–64

NONPARAMETRIC QUASI-LIKELIHOOD[1]

BY JENG-MIN CHIOU AND HANS-GEORG MÜLLER

National Chung Cheng University and University of California, Davis

The quasi-likelihood function proposed by Wedderburn broadened the scope of generalized linear models by specifying the variance function instead of the entire distribution. However, complete specification of variance functions in the quasi-likelihood approach may not be realistic. We define a nonparametric quasi-likelihood by replacing the specified variance function in the conventional quasi-likelihood with a nonparametric variance function estimate. This nonparametric variance function estimate is based on squared residuals from an initial model fit. The rate of convergence of the nonparametric variance function estimator is derived. It is shown that the asymptotic limiting distribution of the vector of regression parameter estimates is the same as for the quasi-likelihood estimates obtained under correct specification of the variance function, thus establishing the asymptotic efficiency of the nonparametric quasi-likelihood estimates. We propose bandwidth selection strategies based on deviance and Pearson's chi-square statistic. It is demonstrated in simulations that for finite samples the proposed nonparametric quasi-likelihood method can improve upon extended quasi-likelihood or pseudo-likelihood methods where the variance function is assumed to fall into a parametric class with unknown parameters. We illustrate the proposed methods with applications to dental data and cherry tree data.

1. Introduction. Generalized linear models (GLMs) were introduced by Nelder and Wedderburn (1972) as a unifying concept. An important extension proposed by Wedderburn (1974) is the quasi-likelihood function, which requires assumptions on the first two moments only, rather than the entire distribution of the data. The quasi-likelihood approach is useful because in many situations the exact distribution of the observations is unknown. Moreover, a quasi-likelihood function has statistical properties similar to those of a log-likelihood function.

Since the variance function is an essential determinant of the quasi-likelihood, its specification is an important problem in the quasi-likelihood approach. In many applications, it is a priori unclear how the variance function should be specified. However, efficient inference for the regression parameters relies on correct variance functions. Nelder and Pregibon (1987) proposed an extended quasi-likelihood function which incorporates the estimation of variance functions in a parametric setting. This parametric approach broadened the use of quasi-likelihood functions by assuming a parametric form of the variance function. Another parametric approach is the pseudo-likelihood

Received September 1997; revised July 1998.

[1] Supported in part by NSF Grant DMS-96-25984 and by NSA Grant MDA-904-96-10026.

AMS 1991 *subject classifications*. 62G07, 62G10, 62J12.

Key words and phrases. Variance function, smoothing, semiparametric modeling, deviance, bandwidth selection, efficiency, extended quasi-likelihood, pseudo-likelihood.

method of variance function estimation introduced by Carroll and Ruppert (1982). These two approaches were compared by Davidian and Carroll (1988) and Nelder and Lee (1992); see also McCullagh and Nelder (1989).

In this study, we extend the quasi-likelihood approach for the situation of a known link function by defining a nonparametric quasi-likelihood suitable for situations where the variance function is unknown but can be assumed to be smooth. The nonparametric quasi-likelihood is obtained by substituting a nonparametrically estimated variance function in lieu of the true variance function in the usual definition of the quasi-likelihood. The nonparametric variance function estimate which is used in the nonparametric quasi-likelihood is then obtained by smoothing squared residuals obtained from a previous model fit which are centered at the estimated means. As a smoothing method, we choose local polynomial fitting by locally weighted least squares, a long established smoothing method which is described in detail in Fan and Gijbels (1996). Other smoothing methods such as kernel smoothers or smoothing splines could be used equally well. The estimates of the regression parameters are then obtained by maximizing this nonparametric quasi-likelihood. Typically, this procedure is iterated by using the updated model parameters in order to obtain new residuals and estimated means and thus an updated nonparametric variance function estimate, which then in turn can be used to obtain improved parameter estimates.

We establish consistency and rate of convergence of the corresponding variance function estimators in Theorem 4.1, and the asymptotic efficiency of the nonparametric quasi-likelihood estimators (NQLEs) for the parameter vector in Theorem 4.2 (Section 4). More details on the iterations and in particular the proposed bandwidth selectors are provided in Section 5. The finite sample behavior of the proposed estimators is investigated in Section 6 by means of simulation studies. This includes comparisons with parametric variance function estimators, which form the core of the extended quasi-likelihood and the pseudo-likelihood methods. We demonstrate that the proposed nonparametric quasi-likelihood (NQL) method leads to improvements upon the parametric quasi-likelihood methods. Section 7 contains illustrative examples where the nonparametric quasi-likelihood method is applied to dental data concerning the relation between force and electrical activity in the chewing muscle and to two-dimensional data on the volume of cherry trees. All proofs and auxiliary results are compiled in Section 8. In Section 2, the main assumptions of the proposed nonparametric quasi-likelihood model are introduced, and the nonparametric variance function estimators are defined in Section 3.

2. Nonparametric quasi-likelihood model.

A nonparametric quasi-likelihood model can be written as follows:

$$Y_i = g(x_i^T \beta) + \varepsilon_i, \tag{M1}$$

where $g(\cdot)$ is denoted as the link function following Weisberg and Welsh (1994). We note that $g(\cdot)$ is often referred to as the inverse link function in the literature on generalized linear models. Furthermore, x_i is the nonrandom

p-dimensional predictor variable corresponding to the ith observation Y_i, β is the p-dimensional vector of regression parameters to be estimated, and the errors ε_i are independent, satisfying $E\varepsilon_i = 0$, $E\varepsilon_i^2 < \infty$ and $\operatorname{var}(\varepsilon_i^2) > 0$.

(M2) There exists a function $\sigma^2(\cdot)$, $\sigma^2(\cdot) \geq \gamma > 0$ for a $\gamma > 0$, such that

$$\operatorname{var}(\varepsilon_i) = \operatorname{var}(Y_i) = \sigma^2\big(g(x_i^T\beta)\big) = \sigma^2(EY_i).$$

The variance of the observations is a function of the mean, referred to as the variance function.

Throughout it is assumed that the link function is given and the variance function is unknown and is to be estimated. Further model assumptions are as follows. A smoothness assumption:

(M3) The link function $g(\cdot)$ is three times and the variance function $\sigma^2(\cdot)$ is twice continuously differentiable with bounded derivatives.

A moment assumption necessary for obtaining uniform consistency of the variance function estimator, as well as bounds for replacing the estimated variance function in the nonparametric quasi-score function:

(M4) There exists a function $\mu_4(\cdot)$ such that $E\varepsilon_i^4 = \mu_4(EY_i)$. The function $\mu_4(\cdot)$ is continuous; furthermore, there exists a $s > 2$ such that $\max_{1\leq i\leq n} E\varepsilon_i^{2s} < c < \infty$ for some $c > 0$.

(M5) There exists a $M > 0$ such that $\max_{1\leq i\leq n} \|x_i\| \leq M < \infty$, for all n.

We assume that the covariates x_i are fixed, and (M5) ensures that the linear predictors are bounded. Furthermore, given the link function g and the parameter vector β, we assume that $\{x_1, x_2, \ldots, x_n\}$ form a sequence of designs such that the means $\mu_i = g(x_i^T\beta)$ are generated by a "design density" f_μ which is assumed to satisfy the following conditions:

(M6) The support of f_μ is a compact interval, $\int f_\mu(u)\,du = 1$, and f_μ is twice continuously differentiable, satisfying $0 < \inf f_\mu(\cdot) \leq \sup f_\mu(\cdot) < \infty$. The design points $\{x_1, x_2, \ldots, x_n\}$ are chosen in such a way that the values $\mu_i = g(x_i^T\beta)$ satisfy

$$\int_{-\infty}^{\mu_i} f_\mu(u)\,du = \frac{i-1}{n-1} \quad \text{for all } n.$$

This also includes discrete or binary predictors in cases where the number of combinations of levels of the predictor variables is large. Assumption (M6) will be needed for asymptotic approximations of sums by integrals.

(M7) There exists a positive definite matrix Σ such that, as $n \to \infty$,

$$\frac{1}{n}(D^T V^{-1} D) \to \Sigma.$$

Here, D is the $n \times p$ matrix of full rank with elements $D_{ir} = g'(\eta_i)x_{i(r-1)}$, $\eta_i = x_i^T\beta$ for $1 \leq i \leq n$ and $1 \leq r \leq p$, setting $x_{i0} = 1$, and V^{-1} is a diagonal matrix with elements $\{\sigma^2(\mu_i)\}^{-1}$ for $1 \leq i \leq n$.

This condition is needed to obtain the asymptotic covariance matrix of the nonparametric quasi-likelihood estimators $\hat{\beta}^*$.

Following Wedderburn's (1974) quasi-likelihood approach with known variance function $\sigma^2(\cdot)$, the quasi-likelihood is

$$Q(\mu, y) = \sum_{i=1}^{n} \int_{y_i}^{\mu_i} \frac{y_i - t}{\sigma^2(t)}\, dt. \tag{2.1}$$

In the case of unknown variance function, we propose to replace the variance function $\sigma^2(\cdot)$ in (2.1) with a nonparametrically estimated variance function estimator $\sigma_n^2(\cdot)$. Then a nonparametric quasi-likelihood is

$$Q^*(\mu, y) = \sum_{i=1}^{n} \int_{y_i}^{\mu_i} \frac{y_i - t}{\sigma_n^2(t)}\, dt. \tag{2.2}$$

The nonparametric quasi-score function $U^*(\beta)$ for this nonparametric quasi-likelihood is then

$$\begin{aligned} U^*(\beta) &= \sum_{i=1}^{n} \frac{y_i - \mu_i}{\sigma_n^2(\mu_i)} g'(\eta_i) x_i \\ &= D^T V_n^{-1}(y - \mu), \end{aligned} \tag{2.3}$$

where $x_i = (1, x_{i1}, x_{i2}, \dots, x_{i(p-1)})^T$, V_n^{-1} is a diagonal matrix with elements $\{\sigma_n^2(\mu_i)\}^{-1}$ for $1 \le i \le n$ and D is defined in (M7). The nonparametric quasi-likelihood estimator (NQLE) $\hat{\beta}^*$ of β is a solution of the estimating equation

$$U^*(\beta) = 0. \tag{2.4}$$

To obtain $\hat{\beta}^*$, the Newton–Raphson method with scoring can be applied iteratively until convergence occurs, see Section 5 for more details. A nonparametric quasi-deviance is then naturally defined as

$$D(y; \mu, \sigma_n^2) = -2 \sum_{i=1}^{n} \int_{y_i}^{\mu_i} \frac{y_i - t}{\sigma_n^2(t)}\, dt. \tag{2.5}$$

A noteworthy feature is that this nonparametric quasi-deviance is "scaled" as compared to the (nonscaled) quasi-deviance in a generalized linear model since the dispersion parameter is absorbed into the nonparametric variance function. Consequently, the nonparametric quasi-deviance, $D(y; \mu, \sigma_n^2)$, can be used as a goodness-of-fit statistic in a straightforward manner with its expected value approximately equal to the degrees of freedom. A "nonparametric" Pearson chi-square statistic, $X^2(y; \mu, \sigma_n^2)$, is defined likewise as

$$X^2(y; \mu, \sigma_n^2) = \sum_{i=1}^{n} \frac{(y_i - \mu_i)^2}{\sigma_n^2(\mu_i)} \tag{2.6}$$

and can be used as a goodness-of-fit statistic, too. Not to have to worry about overdispersion is one of the advantages of the nonparametric quasi-likelihood approach. The case of an unknown link function has been discussed recently in Chiou and Müller (1998).

3. Nonparametric variance function estimation in quasi-likelihood models. A consistent variance function estimator is essential in nonparametric quasi-likelihood; otherwise, the NQLE $\hat{\beta}^*$ will not be efficient and goodness-of-fit statistics and statistical inferences may not be valid. The proposed nonparametric variance function estimator is motivated by the local averaging of squared residuals. Methods for variance function estimation have been proposed based on differencing schemes with the aim of asymptotically eliminating the effect of a smooth mean function [compare Hall and Carroll (1989), Hall, Kay and Titterington (1990) and Müller and Stadtmüller (1987)]. Squared generalized differences are then smoothed to obtain the variance function estimate. An alternative is to smooth squared residuals, which are obtained from a prior nonparametric regression fit [see Silverman (1985), Fan and Gijbels (1996) and Ruppert, Wand, Holst and Hössjer (1997)]. Most of these estimators can be viewed as local quadratic forms, applied to the vectors of data falling into the smoothing window [Müller and Stadtmüller (1993)]. We use here the approach of obtaining squared residuals from a previous fit of the model and then smoothing them by applying the local polynomial smoothing method. The estimation of variance functions by local polynomial fitting of squared residuals obtained from a nonparametric regression fit was studied in detail by Ruppert, Wand, Holst and Hössjer (1997).

For our main results, the choice of a smoothing method is incidental and any reasonable smoother could be used in lieu of local polynomial fitting. Our basic idea is to combine the quasi-likelihood approach in estimating the regression coefficients with nonparametric regression techniques to obtain the variance function estimates.

From (M1) and (M2), $\varepsilon_i^2 = (Y_i - \mu_i)^2$ and $E[\varepsilon_i^2] = \sigma^2(\mu_i)$, and we can write a variance function model as follows:

$$\varepsilon_i^2 = \sigma^2(\mu_i) + \delta_i, \qquad i = 1, 2, \dots, n, \tag{3.1}$$

where δ_i is an error term with $E\delta_i = 0$.

Let I be a compact interval, $I \subset \operatorname{int}\{\operatorname{support}(f_\mu)\}$, where $\operatorname{int}\{A\}$ denotes the interior of a set A, such that there exists a $\rho > 0$ and for any $x \in I$ it holds that $[x - \rho, x + \rho] \subset \operatorname{support}(f_\mu)$. We consider only $u \in I$ to avoid the consideration of boundary effects. [We note that the following results can be extended to cover the case $I = \operatorname{support}(f_\mu)$ which increases the technical and notational burden.] Let

$$\sigma_n^2(u) = \sum_{i=1}^{n} W_{ni}(u)\varepsilon_i^2, \tag{3.2}$$

where

$$W_{ni}(u) = W_{ni}(u; \mu_1, \mu_2, \dots, \mu_n) \tag{3.3}$$

are local linear weight functions. The weight functions $W_{ni}(\cdot)$ are derived from fitting local polynomials by weighted least squares or by using kernel estimators. Other smoothers could be used as well. If we fit local lines and use a nonnegative kernel K as the weight function, we find the following explicit form for the weights $W_{ni}(u)$ [see, e.g., Fan and Gijbels (1996)]:

$$W_{ni}(u) = \frac{(1/nb)K((\mu_i - u)/b)\{A_{n,2}(u) - (\mu_i - u)A_{n,1}(u)\}}{A_{n,0}(u)A_{n,2}(u) - A_{n,1}^2(u)}, \tag{3.4}$$

where

$$A_{n,j}(u) = \frac{1}{nb}\sum_{i=1}^{n} K\left(\frac{\mu_i - u}{b}\right)(\mu_i - u)^j, \qquad j = 0, 1, 2. \tag{3.5}$$

The basic assumptions on the kernel function K in (3.4) and (3.5) are

(K1) support$(K) = [-1, 1]$, $\int_{-1}^{1} K(v)\,dv = 1$, K is continuously differentiable on $[-1, 1]$, Lipschitz$[-1, 1]$, $K(v) = K(-v)$, $K \geq 0$.

Note that (K1) implies that $\int v^j K(v)\,dv = 0$ for odd j, and $\int v^j K(v)\,dv < \infty$ for even j, and also that K is bounded.

The basic requirements for the sequence of bandwidths $b = b(n) > 0$ are

(K2) $b \to 0$, $nb^2 \to \infty$ as $n \to \infty$.

A further assumption on the sequence of bandwidths is needed for uniform convergence results. Assume that for a constant r with $2 < r < s$, where s is as in (M4), it holds that

(K3) $\liminf_{n\to\infty} \left(\frac{nb}{\log n}\right)^{1/2} n^{-2/r} > 0$.

Since the μ_i in (3.2) are actually unknown, the ε_i^2 are not observable. Therefore, we need to replace the μ_i, ε_i^2 and $W_{ni}(u)$ in (3.2) with estimated values $\hat{\mu}_i$, $\hat{\varepsilon}_i^2$ and $\hat{W}_{ni}(u)$, respectively, where

$$\hat{\mu}_i = g(x_i^T \hat{\beta}^*), \tag{3.6}$$

$$\hat{\varepsilon}_i^2 = (Y_i - \hat{\mu}_i)^2 \tag{3.7}$$

and

$$\hat{W}_{ni}(u) = W_{ni}(u; \hat{\mu}_1, \hat{\mu}_2, \ldots, \hat{\mu}_n). \tag{3.8}$$

Here, $\hat{\beta}^*$ is the NQLE of β. Given $\{(\hat{\mu}_i, \hat{\varepsilon}_i), i = 1, 2, \ldots, n\}$, the nonparametric estimators of the variance function are constructed by

$$\hat{\sigma}_n^2(u) = \sum_{i=1}^{n} \hat{W}_{ni}(u)\,\hat{\varepsilon}_i^2, \tag{3.9}$$

where

$$\hat{W}_{ni}(u) = \frac{(1/nb)K((\hat{\mu}_i - u)/b)\{\hat{A}_{n,2}(u) - (\hat{\mu}_i - u)\hat{A}_{n,1}(u)\}}{\hat{A}_{n,0}(u)\hat{A}_{n,2}(u) - \hat{A}_{n,1}^2(u)} \tag{3.10}$$

and

$$\hat{A}_{n,j}(u) = \frac{1}{nb}\sum_{i=1}^{n} K\left(\frac{\hat{\mu}_i - u}{b}\right)(\hat{\mu}_i - u)^j, \qquad j = 0, 1, 2. \tag{3.11}$$

The smoothing step is not difficult because it is a univariate smoothing procedure even for the case of multiple predictor variables. The common "curse-of-dimensionality" in nonparametric regression is therefore not a problem.

4. Asymptotic properties. We aim to show that when the variance function is unknown and is replaced with consistent nonparametric variance function estimates, the $\sqrt{n}$-consistency and the asymptotic normality properties of the NQLE of β are the same as those for the quasi-likelihood estimator of β obtained under a correctly specified variance function. Consistency of the nonparametric variance function estimates plays a central role in obtaining the asymptotic results. Proofs of the following results and required auxiliary results can be found in Section 8. First, in Lemma 4.1, we show that $\sigma_n^2(u)$ as defined in (3.2) is a consistent estimator when the "design points" μ_i, evaluated at the true values of the parameter vector β, are not contaminated with errors. Based on this result, we then proceed to show the consistency of $\hat{\sigma}_n^2(u)$.

LEMMA 4.1. *Under* (M1)–(M6) *and* (K1)–(K3),

$$\sup_{u\in I}\big|E\sigma_n^2(u)-\sigma^2(u)\big| = O(b^2), \tag{4.1}$$

$$\sup_{u\in I} E\big[(\sigma_n^2(u)-\sigma^2(u))^2\big] = O\bigg(\frac{1}{nb}+b^4\bigg), \tag{4.2}$$

$$\sup_{u\in I}\big|\sigma_n^2(u)-\sigma^2(u)\big| = O_p\bigg(\bigg[\frac{\log n}{nb}\bigg]^{1/2}+b^2\bigg), \tag{4.3}$$

where $\sigma_n^2(u)=\sum_{i=1}^n W_{ni}(u)\varepsilon_i^2$ *as defined in* (3.2).

Since the μ_i and the ε_i^2 in (3.2) are unknown and need to be replaced with the $\hat{\mu}_i$ and the $\hat{\varepsilon}_i^2$ as in (3.9), the smoothing scatterplot data $\{(\hat{\mu}_i,\hat{\varepsilon}_i^2)\}$ for obtaining $\hat{\sigma}_n^2(\cdot)$ are contaminated with errors. For general results on nonparametric regression with errors in variables, see, for example, Fan and Truong (1993). The following result demonstrates uniform convergence of $\hat{\sigma}_n^2(u)$ toward $\sigma^2(u)$ in this situation, provided a $\sqrt{n}$-consistent initial estimator for the parameters β is available.

THEOREM 4.1. *Under* (M1)–(M6) *and* (K1)–(K3), *if* $\|\hat{\beta}-\beta\| = O_p(1/\sqrt{n})$, *then, for* $u\in I$,

$$\sup_{u\in I}\big|\hat{\sigma}_n^2(u)-\sigma^2(u)\big| = O_p\bigg(\bigg[\frac{\log n}{nb}\bigg]^{1/2}+b^2+\frac{1}{\sqrt{n}b}\bigg), \tag{4.4}$$

where $\hat{\sigma}_n^2(u)=\sum_{i=1}^n \hat{W}_{ni}(u)\hat{\varepsilon}_i^2$ *as defined in* (3.9).

The main result demonstrating asymptotic efficiency of the NQLE $\hat{\beta}^*$ is the following theorem.

THEOREM 4.2. *In a nonparametric quasi-likelihood model, assume that* (M1)–(M7) *and* (K1)–(K3) *are satisfied, and that the variance function is estimated by* $\hat{\sigma}_n^2(\cdot)$ *in* (3.9). *The estimates* $\hat{\sigma}_n^2(\cdot)$ *are truncated below by a sequence* $\zeta_n > 0$, *where* $\zeta_n \to 0$. *This sequence satisfies*

(K4) $b/\zeta_n \to 0$, $nb^2\zeta_n^2 \to \infty$ *and* $nb\zeta_n^2/\log n \to \infty$.

Then the NQLE $\hat{\beta}^*$ *in* (2.4) *is asymptotically normally distributed such that, as* $n \to \infty$,

$$\sqrt{n}(\hat{\beta}^* - \beta) \to_D N_p(\mathbf{0}, \Sigma^{-1}), \tag{4.5}$$

where $\Sigma = \lim_{n\to\infty}(1/n)D^T V^{-1} D$, *as defined in* (M7).

Note that (4.5) implies that the NQLE $\hat{\beta}^*$ has the same asymptotic distribution as the QLE obtained with known variance function [see McCullagh (1983)]. A practical way to utilize the distribution of the NQLEs is obtained as a consequence of Theorem 4.2.

COROLLARY 4.1. *Under the assumptions of Theorem* 4.2, *let* $\hat{\beta}^*$ *be the NQLE* (2.4) *and let*

$$\hat{\Sigma}^{-1} = n\,(\hat{D}^T \hat{V}^{-1} \hat{D})^{-1}, \tag{4.6}$$

where $\hat{D} = (\hat{D}_{ir})_{1\le i\le n,\, 1\le r\le p}$ *with* $\hat{D}_{ir} = g'(x_i^T \hat{\beta}^*)\, x_{i(r-1)}$, *setting* $x_{i0} = 1$, *and* $\hat{V}^{-1} = \operatorname{diag}(\{\hat{\sigma}_n^2(\hat{\mu}_i)\}^{-1})_{1\le i\le n}$ *with* $\hat{\mu}_i = g(x_i^T \hat{\beta}^*)$. *Then,*

$$\hat{\Sigma}^{-1} \to_p \Sigma^{-1}, \tag{4.7}$$

$$\hat{\beta}^* \sim N_p(\beta, (\hat{D}^T \hat{V}^{-1} \hat{D})^{-1}), \tag{4.8}$$

for large n.

Based on the limiting distribution and the covariance matrix of the NQLE $\hat{\beta}^*$, we can develop an asymptotic test for the class of hypotheses:

$$H_0\colon \Lambda\beta = \zeta_0 \quad \text{versus} \quad H_{1n}\colon \Lambda\beta = \zeta_{1n}, \tag{4.9}$$

where Λ is a $m \times p$ matrix of rank m, and ζ_0 and ζ_{1n} are $m \times 1$ vectors, $m \le p$. Consider the test statistics

$$T_n = n(\Lambda\hat{\beta}^* - \zeta_0)^T(\Lambda\hat{\Sigma}^{-1}\Lambda^T)^{-1}(\Lambda\hat{\beta}^* - \zeta_0). \tag{4.10}$$

Note that under the null hypothesis H_0, $T_n \to_D \chi_m^2$ where χ_m^2 has a central χ^2 distribution with m degrees of freedom. On the other hand, under the alternatives H_{1n}, $T_n \to_D \chi_m^2(\rho^2)$ where $\chi_m^2(\rho^2)$ has a noncentral χ^2 distribution with m degrees of freedom and noncentrality parameter ρ^2 by assuming that

$$n(\zeta_{1n} - \zeta_0)^T(\Lambda\hat{\Sigma}^{-1}\Lambda^T)^{-1}(\zeta_{1n} - \zeta_0) \to \rho^2,$$

for a fixed real constant ρ. The null hypothesis H_0 in (4.9) is rejected at level α if $T_n > \chi_{m;\alpha}^2$, where $\chi_{m;\alpha}^2$ is the $100(1-\alpha)\%$ quantile of the corresponding χ^2

distribution with m degrees of freedom. Moreover, a $100(1-\alpha)\%$ confidence region for β is given by $\{\beta\colon n(\hat{\beta}^*-\beta)^T\hat{\Sigma}(\hat{\beta}^*-\beta)\le \chi^2_{p;1-\alpha}\}$.

5. Iterative estimation and automatic bandwidth selection.

For the practical implementation of NQLEs, we propose an iterative updating procedure for estimating the regression parameter β and the variance function $\sigma^2(\cdot)$. The updating steps alternate between a Newton–Raphson scoring step and a smoothing step. We introduce a generic notation S for smoothing scatterplot data $(\hat{\mu}_i, \tilde{\sigma}_i^2)$ based on weight functions $\hat{W}_{ni}(u)$ such that

$$S\big(u, b; (\hat{\mu}_i, \tilde{\sigma}_i^2), i=1,2,\dots,n\big) = \sum_{i=1}^{n} \hat{W}_{ni}(u)\tilde{\sigma}_i^2, \tag{5.1}$$

where b is the bandwidth, $\hat{W}_{ni}(u) = W_{ni}(u; \hat{\mu}_1, \hat{\mu}_2, \dots, \hat{\mu}_n)$ as in (3.8) and (3.10), (3.11), $\hat{\mu}_i = x_i^T\hat{\beta}$ and $\tilde{\sigma}_i^2 = (y_i-\hat{\mu}_i)^2$.

The two-stage iterative estimating procedure can be summarized as follows.

1. For an assumed variance function $\sigma^2(\cdot)$, the parameter estimates $\hat{\beta}^*$ are obtained by the standard quasi-likelihood approach. For the initial step, we can simply use constant variance functions $\sigma^2(\cdot)\equiv 1$. This will lead to $\sqrt{n}$-consistent parameter estimates.
2. For fixed β, nonparametric variance function estimates $\hat{\sigma}_n^2$ are obtained by $\hat{\sigma}_n^2(u) = S(u, b; (\hat{\mu}_i, \tilde{\sigma}_i^2), i=1,2,\dots,n)$ as defined in (5.1).

Based on the two-stage iterative estimating procedure, β is first estimated pretending that $\sigma^2(\cdot)$ is known, then $\sigma^2(\cdot)$ is estimated pretending β is known. This iterative procedure is continued until some convergence criterion is met. We note that step 1 corresponds to maximizing a likelihood while step 2 is a univariate smoothing step. This smoothing step requires a bandwidth choice, which generally is known to determine to a large extent the quality of the estimated curve. Without a carefully chosen bandwidth, good estimates of the variance function cannot be obtained: goodness-of-fit statistics may be misleading, and the parameter estimates may be adversely affected. It is possible to use conventional bandwidth selection schemes, such as cross-validation and plug-in methods in this context.

We propose a new bandwidth selection method based on the goodness-of-fit statistics which leads to good practical results. The motivation of the proposed method derives from the asymptotic results that both the expected value of nonparametric quasi-deviance, $E(D)$, and Pearson's chi-square statistic, $E(X^2)$, are approximately equal to the degrees of freedom, $n-p$. Since both D and X^2 depend on the estimated variance, this equality can be utilized for bandwidth choice as follows.

Let $\hat{\mu}_b$ denote the estimated value of μ where the nonparametric variance function estimates are smoothed with the bandwidth b. Define

$$G_D\big(b; \hat{\mu}_b, \hat{\sigma}_n^2(\hat{\mu}_b)\big) = \big|D(y; \hat{\mu}_b, \hat{\sigma}_n^2(\hat{\mu}_b)) - (n-p)\big|, \tag{5.2}$$

where D is defined in (2.5), as a measure of the discrepancy between the nonparametric quasi-deviance and its approximated expected value, where the estimates $\hat{\mu}_b$ and $\hat{\sigma}_n^2(\hat{\mu}_b)$ both depend on the bandwidth b. The “optimal” bandwidth is then chosen as

$$\hat{b}_{\text{opt}, D} = \arg\min_b G_D\big(b; \hat{\mu}_b, \hat{\sigma}_n^2(\hat{\mu}_b)\big). \tag{5.3}$$

Analogously, the bandwidth choice criterion can be based on Pearson's X^2 (2.6) with

$$\hat{b}_{\text{opt}, P} = \arg\min_b G_P\big(b; \hat{\mu}_b, \hat{\sigma}_n^2(\hat{\mu}_b)\big), \tag{5.4}$$

where

$$G_P\big(b; \hat{\mu}_b, \hat{\sigma}_n^2(\hat{\mu}_b)\big) = \big|X^2(y; \hat{\mu}_b, \hat{\sigma}_n^2(\hat{\mu}_b)) - (n-p)\big|. \tag{5.5}$$

We use newly estimated optimal bandwidths for each iteration where the variance function is updated.

The proposed methods and in particular bandwidth choices are also of potential interest for image smoothing. In the context of multinomial smoothing, Titterington (1985), Section 3.4, has suggested bandwidth choices which are closely related to (5.5). These bandwidths minimize the difference between the Pearson's χ^2 distance of the data and the smoothed estimates and the expected value of this Pearson's χ^2. This bandwidth selection method was further explored in Thompson, Brown, Kay and Titterington (1991).

6. Simulation studies.

6. Simulation studies. Simulation studies were performed to investigate the effect of variance function estimation in quasi-likelihood models. The underlying data distributions were overdispersed Poisson and Binomial. Four methods were compared in the simulation study, namely (a) the standard quasi-likelihood (QL) method where the variance function is assumed known, the best possible method, which serves as a gold standard; (b) the extended quasi-likelihood (EQL) method, with the common power-of-the-mean variance function, $\text{var}(y_i) = \phi(Ey_i)^\lambda$ (correctly parameterized for overdispersed Poisson but incorrect for Binomial); (c) the pseudo-likelihood (PL) method, also with power-of-the-mean variance function; (d) the nonparametric quasi-likelihood (NQL) method. We implemented both deviance and Pearson based bandwidth selectors (5.3) and (5.4) and found that they perform very similarly with a slight advantage for deviance-based bandwidth selection. Therefore, the results reported here use this selector. Thus, the implementation of NQL as studied in this simulation is fully automatic and does not require a bandwidth choice by the user.

For each iteration, 1000 Monte Carlo runs were made with a univariate predictor variable. We generated overdispersed Poisson data via a Gamma–Poisson mixture. The link function $\mu_i = e^{\eta_i}$ was chosen, $\eta_i = \beta_0 + \beta_1 x_i$, with $\beta_0 = 3$, $\beta_1 = 4$, and the dispersion parameter was $\phi = 5$, so that $\text{var}(y_i) = 5\mu_i$. The sample size was $n = 100$ with design points $x_i = i/n$, $i = 1, \dots, 100$. The results for the parameter estimates $\hat{\beta}_0$ and $\hat{\beta}_1$ are in Table 1. The gold

TABLE 1
Simulation results of the estimated regression parameters ($\beta_0 = 3$, $\beta_1 = 4$) for overdispersed Poisson data

Parameter	**Method**[*]	**Estimated mean**	**Sample std. err.**	**Bias**	**MSE**	**Relative**[†] **MSE**
$\hat{\beta}_0$	QL	3.0012	0.0519	0.0012	0.002694	1.0000
	EQL	3.0001	0.0522	0.0001	0.002722	1.0104
	PL	2.9994	0.0523	−0.0006	0.002736	1.0156
	NQL	2.9998	0.0549	−0.0002	0.003009	1.1169
$\hat{\beta}_1$	QL	3.9989	0.0647	−0.0011	0.004182	1.0000
	EQL	4.0002	0.0653	0.0002	0.004262	1.0191
	PL	4.0013	0.0651	0.0013	0.004245	1.0151
	NQL	3.9999	0.0701	−0.0001	0.004909	1.1738

[*] QL: quasi-likelihood, EQL: extended quasi-likelihood. PL: pseudo-likelihood. NQL: nonparametric quasi-likelihood.
[†] Relative performance as compared to QL, measured by the ratio of MSE to QL.

standard is the QL method with a correct variance function. The EQL and PL methods do predictably well here since they were implemented with the power-of-the-mean variance function, which happens to be the correct type of variance function. The NQL method does less well here as it hedges against a host of other possible smooth alternatives.

Another criterion is how well the asymptotic approximations made for inference are justified for the various methods. The empirical coverage frequencies and average lengths for 90% and 95% asymptotic confidence intervals for single parameters constructed via Corollary 4.1 are shown in Table 2. The performance of the nonparametric method is seen to be surprisingly good both in terms of coverage probabilities and lengths, given that the other methods have the advantage of operating with true or correctly parameterized variance function.

TABLE 2
Coverage of confidence intervals for regression parameters for overdispersed Poisson data

		90% asymptotic confidence			**95% asymptotic confidence**		
Parameter	**Method**	**Percent at miss-left**	**Percent at miss-right**	**length**	**Percent at miss-left**	**Percent at miss-right**	**length**
$\hat{\beta}_0$	QL	4.60	4.40	0.1708	2.30	2.10	0.2034
	EQL	4.40	4.50	0.1734	2.30	2.20	0.2066
	PL	4.40	4.60	0.1740	2.30	2.50	0.2074
	NQL	4.30	4.60	0.1668	2.30	3.10	0.1987
$\hat{\beta}_1$	QL	4.60	4.90	0.2132	1.90	1.80	0.2540
	EQL	4.80	4.80	0.2175	1.80	1.80	0.2591
	PL	4.90	4.80	0.2178	2.00	2.00	0.2595
	NQL	4.60	4.50	0.2043	2.30	2.10	0.2434

TABLE 3
Simulation results of the estimated regression parameters ($\beta_0 = 4$, $\beta_1 = -6$) for Binomial data

Parameter	**Method**	**Estimated mean**	**Sample std. err.**	**Bias**	**MSE**	**Relative MSE**
$\hat{\beta}_0$	QL	3.9796	0.1281	−0.0204	0.016837	1.0000
	EQL	3.9792	0.1501	−0.0208	0.022960	1.3637
	PL	3.9791	0.1501	−0.0209	0.022965	1.3640
	NQL	3.9807	0.1374	−0.0193	0.019245	1.1430
$\hat{\beta}_1$	QL	−5.9666	0.1678	0.0334	0.029272	1.0000
	EQL	−5.9661	0.1943	0.0339	0.038903	1.3290
	PL	−5.9659	0.1943	0.0341	0.038915	1.3294
	NQL	−5.9678	0.1793	0.0322	0.033171	1.1332

As a second example we consider the case of binomial data where the power-of-the-mean variance function model does not apply. Here the underlying variance function is $\mathrm{var}(y_i) = \mu_i(1-\mu_i)$, and the power-of-the-mean versions of EQL and PL are expected to do less well. The link function was chosen as a logit link $\mu_i = e^{\eta_i}/(1+e^{\eta_i})$, $\eta_i = \beta_0 + \beta_1 x_i$, $\beta_0 = 4$, $\beta_1 = -6$, dispersion parameter $\phi = 1$, sample size $n = 100$ with design points

$$x_i = (0.90 - 0.0025i)\mathbf{1}_{\{1\le i\le 80\}} + \big(0.70 - 0.035(i-80)\big)\mathbf{1}_{\{81\le i\le 100\}}.$$

The results for the parameter estimates are in Table 3 (analogous to Table 1). Again, QL serves as a gold standard since it uses the correct variance function. In this setting, EQL and PL methods perform noticeably worse than the NQL method, showing clearly the limitations of these parametric variance function model approaches. The behavior in terms of confidence intervals is comparable among the methods (see Table 4) but NQL again emerges as the winner in terms of lengths of confidence intervals.

TABLE 4
Coverage of confidence intervals for regression parameters for Binomial data

		90% asymptotic confidence			**95% asymptotic confidence**		
Parameter	**Method**	**Percent at miss-left**	**Percent at miss-right**	**length**	**Percent at miss-left**	**Percent at miss-right**	**length**
$\hat{\beta}_0$	QL	3.60	6.50	0.4395	1.60	3.70	0.5237
	EQL	4.10	5.90	0.5359	2.00	3.30	0.6386
	PL	4.10	5.90	0.5356	2.00	3.30	0.6382
	NQL	3.60	7.00	0.4159	2.10	3.30	0.4955
$\hat{\beta}_1$	QL	6.80	2.80	0.5754	3.80	1.60	0.6856
	EQL	6.00	3.90	0.6958	3.60	1.90	0.8291
	PL	6.00	3.90	0.6955	3.60	1.90	0.8287
	NQL	6.90	3.20	0.5511	3.90	1.80	0.6566

7. Examples. As a first example, we consider data from a sample of $n = 150$ subjects who were asked to adjust the force exercised by the chewing muscle in such a way as to achieve a fixed given force as measured by an intraoral probe. Then the intramuscular voltage generated during the muscular contraction constitutes the measurement of the dependent variable. So voltage is a function of force. The relationship is clearly nonlinear, as can been seen from the scatterplot in Figure 1. These data were previously discussed in Müller (1988), pages 36–38.

The link $\mu = \eta^{2/3}$ was assumed from previous investigations. Applying the NQL iteration algorithm to these data then led to the estimated linear predictor

$$\hat{\beta}_0 + \hat{\beta}_1 x = -1948 + 14x,$$

where $\hat{\beta}_0$ has a standard error of 175.8, and $\hat{\beta}_1$ has a standard error of 0.221. Bandwidth selection was based on the deviance criterion (5.3). Figure 2 shows that the estimated variance function is strictly monotone increasing. The shape of the variance function indicates that a power-of-the-mean variance function model does not appear to be adequate for these data. The data fit obtained with the NQL method is shown as a solid curve in Figure 1.

As a second example, we consider data on a sample of felled black cherry trees with measurements of diameter in inches, height in inches and volume in cubic feet [Ryan, Joiner and Ryan (1985)]. The diameters were measured 4 ft. 6 in. above ground level. The purpose of collecting the measurements on the felled trees was to provide a way of predicting the volume of timber

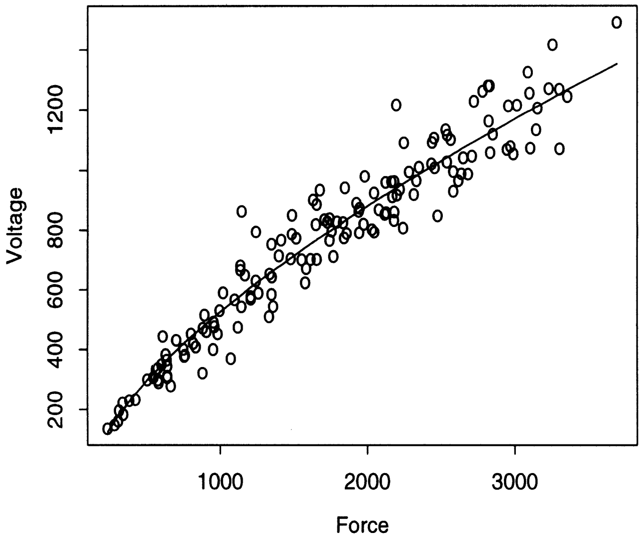


FIG. 1. *Nonparametric quasi-likelihood model with link $g(\eta) = \eta^{2/3}$ for dental data. The scatterplot is observed Voltage versus Force, and the fitted curve corresponds to NQL model fits.*

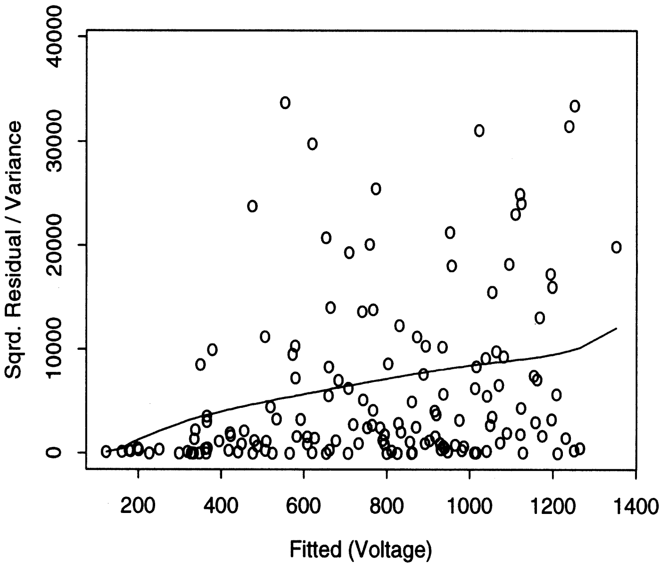


FIG. 2. *Nonparametric quasi-likelihood model with link $g(\eta) = \eta^{2/3}$ for dental data. Fitted curve corresponds to nonparametric variance function estimates based on local averages of squared residuals.*

from their height and diameter measurements. This data set was discussed by Atkinson (1987). One suggestion was to take a power($\frac{1}{3}$) transformation on the response variable. We compare this suggestion of a multiple linear regression where we include a transformed response variable with the nonparametric quasi-likelihood approach. The parameter estimates obtained with the transformation approach are shown in Table 5, and the corresponding fits in Figure 3. The coefficient of determination is $R^2 = 0.9777$ which is very high.

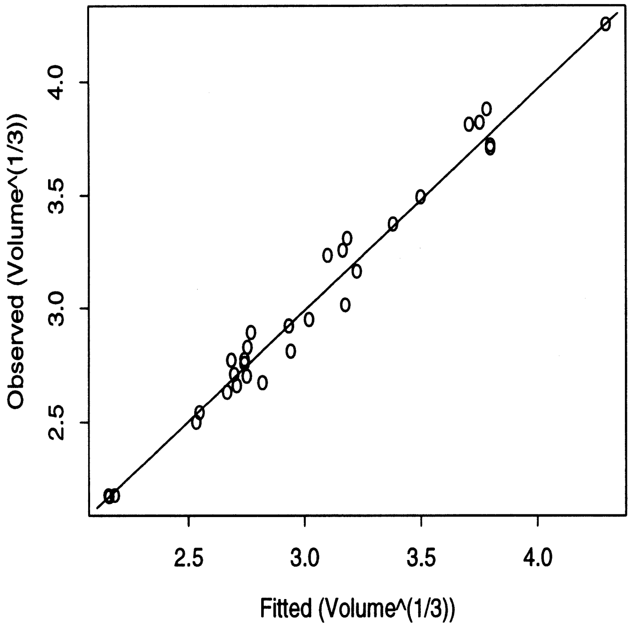
Figure 3 indicates that the power($\frac{1}{3}$) transformation on the response variable fits the data well, overall. We examine the assumption of a constant variance function after using the transformation by applying the nonparametric quasi-likelihood approach with the link function $g(\eta) = \eta^3$ which corresponds

TABLE 5

Estimated regression coefficients for cherry tree data by multiple linear regression model with transformed response volume, by the power($\frac{1}{3}$) transformation

Variable	Estimate	Std. err.	Chi-Square	Pr > Chi
(Intercept)	−0.0854	0.1843	0.21	0.6468
Diameter	0.1515	0.0056	731.90	0.0000
Height	0.0145	0.0028	21.82	0.0000

Note. MSE $= 0.08283$ ($d.f. = 28$), $R^2 = 0.9777$.



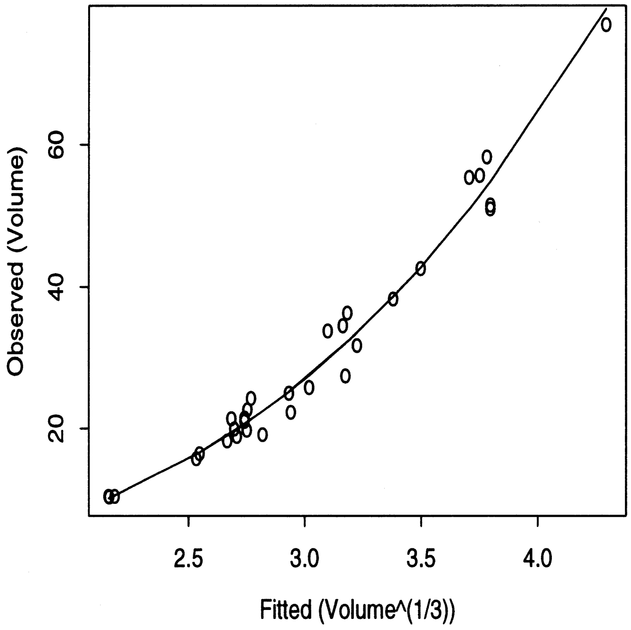


FIG. 3. *Multiple linear regression model with transformed response* volume$^{\frac{1}{3}}$ *for cherry tree data. Upper: Observed values of volume*$^{\frac{1}{3}}$ *versus fitted values of volume*$^{\frac{1}{3}}$*. Lower: Observed values of volume versus fitted values of volume*$^{\frac{1}{3}}$*.*

TABLE 6

Regression coefficients for cherry tree data: nonparametric quasi-likelihood model with link $g(\eta) = \eta^3$ (deviance based bandwidth selection)

Variable	Estimate	Std. err.	Chi-Square	Pr > Chi
(Intercept)	−0.0511	0.1364	0.14	0.7083
Diameter	0.1513	0.0051	880.11	0.0000
Height	0.0141	0.0024	34.52	0.0000

Note. $D = 27.95$, $X^2 = 27.86$, $(d.f. = 28)$, $MSE = 6.5844$. Bandwidth for variance = 64.57 based on deviance.

to the power($\frac{1}{3}$) transformation of the response volume in the multiple linear regression. The results are presented in Table 6, and Figures 4 and 5.

The variance function in Figure 5 is monotone increasing. As diameter and/or the height increase linearly, the prediction of volume is less precise because of the increasing variance. In short, it is harder to predict the volumes of larger trees, and confidence intervals and other inference procedures should be adapted accordingly.

8. Proofs and auxiliary results.

Define $F_\mu(t) = \int_0^t f_\mu(z)\,dz$. Then F_μ^{-1} exists and, according to (M6),

$$\mu_i = F_\mu^{-1}\left(\frac{i-1}{n-1}\right). \tag{8.1}$$

Let

$$\nu_n = \nu_n(u) = \sum_{i=1}^n 1_{\{|\mu_i - u| \le b\}}, \tag{8.2}$$

where $b = b(n) > 0$ is a sequence of bandwidths. Then $\nu_n(u)$ denotes the number of "observations" centered around u within the bandwidth b. It follows from the assumptions on the design density in (M6) that there are constants $0 < C_1 < C_2 < \infty$ such that for sufficiently large n,

$$C_1 \le \frac{\nu_n}{nb} \le C_2 \tag{8.3}$$

uniformly in $u \in I$.

LEMMA 8.1. *For $u \in I$, let $A_{n,j}(u) = (1/nb)\sum_{i=1}^n K((\mu_i - u)/b)(\mu_i - u)^j$ as defined in* (3.5). *Let $\alpha_j = \int v^j K(v)\,dv < \infty$. Note $\alpha_0 = 1$ by* (K1). *Under* (M6) *and* (K1), (K2),

$$A_{n,j}(u) = \begin{cases} b^j \alpha_j f_\mu(u) + O(b^{j+2}) + O\left(\dfrac{b^{j-1}}{n}\right), & \text{for } j = 0, 2, 4, \ldots, \\ b^{j+1}\alpha_{j+1} f'_\mu(u) + O(b^{j+2}) + O\left(\dfrac{b^{j-1}}{n}\right), & \text{for } j = 1, 3, 5, \ldots. \end{cases}$$

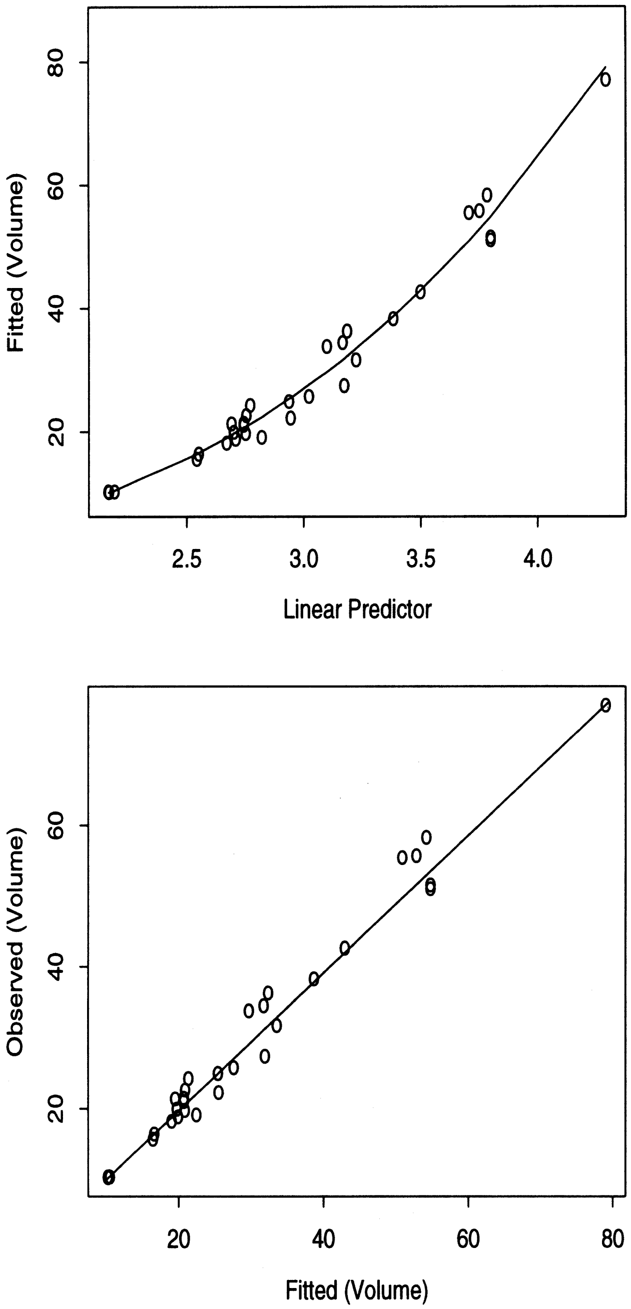


FIG. 4. *Nonparametric quasi-likelihood model with link* $g(\eta) = \eta^3$ *for cherry tree data (bandwidth selection based on nonparametric quasi-deviance). Upper: Fitted curve corresponds to NQL model fits. Lower: Observed versus fitted values of volume.*

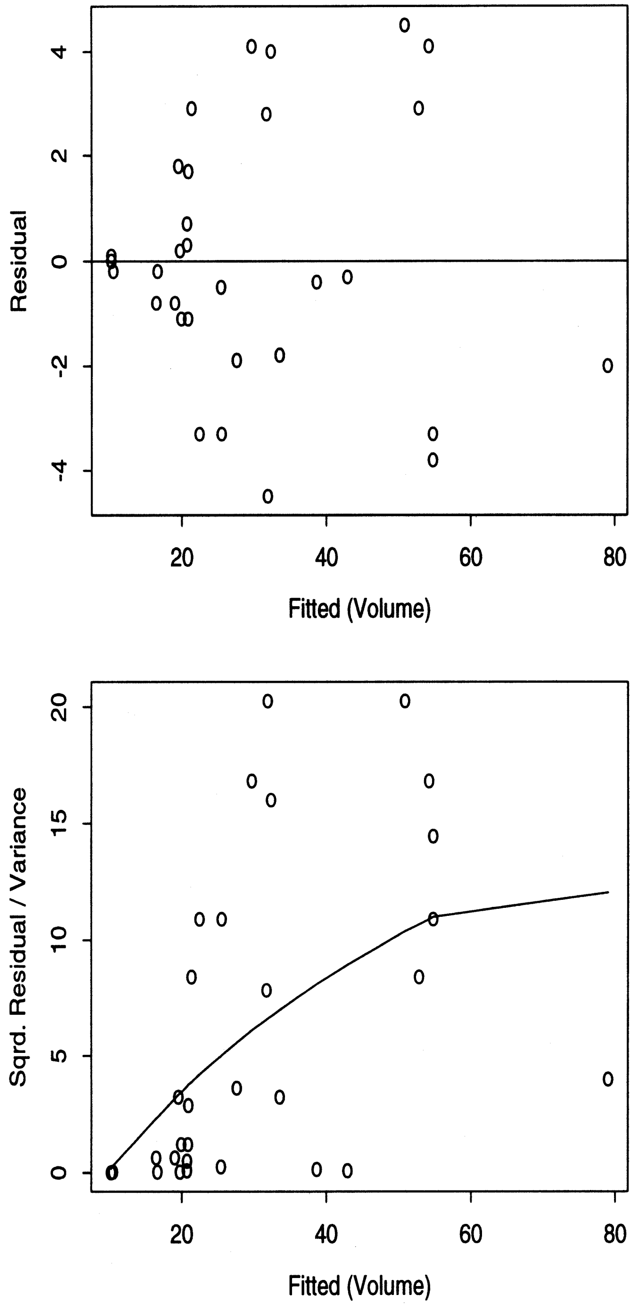


FIG. 5. *Nonparametric quasi-likelihood model with link* $g(\eta) = \eta^3$ *for cherry tree data (bandwidth selection based on nonparametric quasi-deviance). Upper: Residual plot. Lower: Curve corresponds to smoothed variance function estimates based on local averages of squared residuals.*

PROOF. Note $\max_{1\le i\le n}|\mu_i-\mu_{i-1}|=O(1/n)$ under the assumption (M6). By a first-order Taylor series expansion, for all n,

$$\sup_{u\in I}\max_{1<i\le n}\left|K\left(\frac{\mu_i-u}{b}\right)-K\left(\frac{\mu_{i-1}-u}{b}\right)\right|=O\left(\frac{1}{nb}\right).$$

Furthermore, for any $u\in I$ and $1<i\le n$,

$$K\left(\frac{\mu_i-u}{b}\right)(\mu_i-u)^j-K\left(\frac{\mu_{i-1}-u}{b}\right)(\mu_{i-1}-u)^j=O\left(\frac{b^{j-1}}{n}\right).$$

Let $v=(\gamma-u)/b$ and $z=F_\mu(\gamma)$. Then, $dz=bf_\mu(u+bv)\,dv$. Using the above result for Riemann sum approximation and (K1), (K2), it follows that

$$\begin{aligned}A_{n,j}(u)&=\frac{1}{b}\sum_{i=1}^n\frac{1}{n}K\left(\frac{F_\mu^{-1}((i-1)/(n-1))-u}{b}\right)\left(F_\mu^{-1}\left(\frac{i-1}{n-1}\right)-u\right)^j\\&=\int_0^1\frac{1}{b}K\left(\frac{F_\mu^{-1}(z)-u}{b}\right)\left(F_\mu^{-1}(z)-u\right)^j\,dz+\frac{1}{b}\,O(nb)\,O\left(\frac{b^{j-1}}{n^2}\right),\end{aligned}$$

whence the result follows. □

LEMMA 8.2. *For $u\in I$ and $1\le i\le n$, let*

$$\Gamma_{ni}(u)=A_{n,2}(u)-(\mu_i-u)1_{\{|\mu_i-u|\le b\}}A_{n,1}(u),\tag{8.4}$$

$$\Delta_n(u)=A_{n,0}(u)A_{n,2}(u)-A_{n,1}^2(u),\tag{8.5}$$

where $A_{n,j}(u)$ are defined in (3.5). *Then, under* (M6) *and* (K1), (K2),

$$\Gamma_{ni}(u)=b^2\alpha_2f_\mu(u)+O(b^3),\tag{8.6}$$

$$\Delta_n(u)=b^2\alpha_2f_\mu^2(u)+O(b^4)+O\left(\frac{b}{n}\right).\tag{8.7}$$

The proof follows immediately from Lemma 8.1.

LEMMA 8.3. *For $u\in I$ and $1\le i\le n$, $W_{ni}(u)$ is defined in* (3.4). *Then, under* (M6) *and* (K1), (K2),

$$\sum_{i=1}^n W_{ni}(u)(\mu_i-u)^q=\begin{cases}1, & \textit{for } q=0,\\ 0, & \textit{for } q=1,\end{cases}\tag{8.8}$$

$$W_{ni}(u)=\frac{1}{nb}K\left(\frac{\mu_i-u}{b}\right)\Big/f_\mu(u)+O\left(\frac{1}{n}\right).\tag{8.9}$$

PROOF. The first property is referred to as the discrete moment conditions. These can be easily verified by replacing $W_{ni}(u)$ with the explicit form in (3.4). The second property follows from Lemma 8.2. □

LEMMA 8.4. *For $u \in I$, let $A_{n,j}(u)$ and $\hat{A}_{n,j}(u)$ be as defined in* (3.5) *and* (3.11). *Under* (K1) *and* (K2), *if* $\max_{1\le i\le n} |\hat{\mu}_i - \mu_i| = O_p(1/\sqrt{n})$, *then*

$$\hat{A}_{n,j}(u) = A_{n,j}(u) + O_p\left(\frac{b^{j-1}}{\sqrt{n}}\right). \tag{8.10}$$

For the proof, apply (K1), (K2) and the mean value theorem.

LEMMA 8.5. *For any $u \in I$ and $1 \le i \le n$, let*

$$\hat{\Gamma}_{ni}(u) = \hat{A}_{n,2}(u) - (\hat{\mu}_i - u)1_{\{|\hat{\mu}_i - u|\le b\}}\hat{A}_{n,1}(u), \tag{8.11}$$

$$\hat{\Delta}_n(u) = \hat{A}_{n,0}(u)\hat{A}_{n,2}(u) - \hat{A}^2_{n,1}(u), \tag{8.12}$$

where $\hat{A}_{n,j}(u)$ are defined in (3.11). *Under* (K1) *and* (K2), *if* $\max_{1\le i\le n} |\hat{\mu}_i - \mu_i| = O_p(1/\sqrt{n})$, *then*

$$\hat{\Gamma}_{ni}(u) = \Gamma_{ni}(u) + O\left(\frac{b}{\sqrt{n}}\right), \tag{8.13}$$

$$\hat{\Delta}_n(u) = \Delta_n(u) + O\left(\frac{b}{\sqrt{n}}\right). \tag{8.14}$$

PROOF. The result follows since for any $u \in I$ and $1 \le i \le n$,

$$\begin{aligned}\hat{\Gamma}_{ni}(u) = {} & \left(A_{n,2}(u) + O_p\left(\frac{b}{\sqrt{n}}\right)\right) \\ & - \left((\mu_i - u)1_{\{|\mu_i - u|\le b\}} + O_p\left(\frac{1}{\sqrt{n}}\right)\right)\left(A_{n,1}(u) + O_p\left(\frac{1}{\sqrt{n}}\right)\right)\end{aligned}$$

and

$$\begin{aligned}\hat{\Delta}_n(u) = {} & \left(A_{n,0}(u) + O_p\left(\frac{1}{\sqrt{n}b}\right)\right)\left(A_{n,2}(u) + O_p\left(\frac{b}{\sqrt{n}}\right)\right) \\ & - \left(A_{n,1}(u) + O_p\left(\frac{1}{\sqrt{n}}\right)\right)^2. \qquad \square\end{aligned}$$

LEMMA 8.6. *Under* (M1)–(M4), *if $\hat{\beta}$ is a $\sqrt{n}$-consistent estimator of β in the sense that*

$$\|\hat{\beta} - \beta\| = O_p\left(\frac{1}{\sqrt{n}}\right),$$

then

$$\max_{1\le i\le n} |\hat{\mu}_i - \mu_i| = O_p\left(\frac{1}{\sqrt{n}}\right), \tag{8.15}$$

where $\hat{\mu}_i = g(x_i^T\hat{\beta})$ for $i = 1, 2, \ldots, n$.

PROOF. Let $\eta_i = x_i^T\beta$. Then, by uniform boundedness of x_i, $|\hat{\eta}_i - \eta_i| = O_p(n^{-1/2})$ uniformly in i, $1 \le i \le n$. The result follows from (M4). □

LEMMA 8.7. *Let $W_{ni}(\cdot)$ and $\hat{W}_{ni}(\cdot)$ be the weight functions defined in* (3.4) *and* (3.10). *Under* (M6) *and* (K1), (K2), *if $\|\hat{\beta} - \beta\| = O_p(1/\sqrt{n})$, then*

$$\sup_{u\in I}\max_{1\le i\le n}\left|\hat{W}_{ni}(u) - W_{ni}(u)\right| = O_p\left(\frac{1}{nb^2\sqrt{n}}\right). \tag{8.16}$$

PROOF. For any $u \in I$ and $1 \le i \le n$, let

$$\Lambda_{ni}(u) = K\left(\frac{\mu_i - u}{b}\right)\Gamma_{ni}(u),$$

$$\hat{\Lambda}_{ni}(u) = K\left(\frac{\hat{\mu}_i - u}{b}\right)\hat{\Gamma}_{ni}(u),$$

where $\Gamma_{ni}(u)$ and $\hat{\Gamma}_{ni}(u)$ are as in (8.4) and (8.11). Then, for any $u \in I$ and $1 \le i \le n$, using Lemmas 8.5 and 8.6,

$$\begin{aligned}
&\sup_{u\in I}\max_{1\le i\le n}\left|\hat{\Lambda}_{ni}(u) - \Lambda_{ni}(u)\right| \\
&\quad= \sup_{u\in I}\max_{1\le i\le n}\left|\left(K\left(\frac{\mu_i - u}{b}\right) + O_p\left(\frac{1}{\sqrt{n}\,b}\right)\right)\left(\Gamma_{ni}(u) + O_p\left(\frac{b}{\sqrt{n}}\right)\right)\right. \\
&\qquad\qquad\qquad\qquad\qquad\qquad\left. - K\left(\frac{\mu_i - u}{b}\right)\Gamma_{ni}(u)\right| \\
&\quad= O_p\left(\frac{b}{\sqrt{n}}\right).
\end{aligned}$$

Using $\Delta_n(u)$ and $\hat{\Delta}_n(u)$ defined in (8.5) and (8.12), it follows that

$$\begin{aligned}
&\sup_{u\in I}\max_{1\le i\le n}\left|\hat{W}_{ni}(u) - W_{ni}(u)\right| \\
&\quad= \sup_{u\in I}\max_{1\le i\le n}\frac{1}{nb}\left|\frac{\hat{\Lambda}_{ni}(u)}{\hat{\Delta}_n(u)} - \frac{\Lambda_{ni}(u)}{\Delta_n(u)}\right| \\
&\quad= \sup_{u\in I}\max_{1\le i\le n}\frac{1}{nb}\left|\frac{O_p(b/\sqrt{n})}{\Delta_n(u) + O_p(b/\sqrt{n})} - \frac{\Lambda_{ni}(u)O_p(b/\sqrt{n})}{\Delta_n^2(u) + \Delta_n(u)O_p(b/\sqrt{n})}\right| \\
&\quad= O_p\left(\frac{1}{nb^2\sqrt{n}}\right).
\end{aligned}$$
□

LEMMA 8.8. *Under* (M1)–(M6),

$$\sup_{u\in I}\frac{1}{nb}\sum_{i=1}^{n}\varepsilon_i^2 1_{\{|\mu_i - u|\le b\}} = O_p\left(\left[\frac{\log n}{nb}\right]^{1/2}\right) + O(1), \tag{8.17}$$

where $\mu_i = g(x_i^T\beta)$.

PROOF. Let $w_i(u) = (1/nb)1_{\{|\mu_i - u| \le b\}}$. Then

$$\sum_{i=1}^n w_i(u)\varepsilon_i^2 \le \left|\sum_{i=1}^n w_i(u)\big(\varepsilon_i^2 - \sigma^2(\mu_i)\big)\right| + \sum_{i=1}^n \big|w_i(u)\big|\sigma^2(\mu_i).$$

By the boundedness condition of $\sigma^2(\cdot)$, the second term on the right-hand side is $O(1)$. For the first term, we proceed as in Lemma 5.2 in Müller and Stadtmüller (1987), noting that $E\varepsilon_i^2 = \sigma^2(\mu_i)$. □

PROOF OF LEMMA 4.1. For the bias part, by (M4) and with suitable mean values ζ_i, for any $u \in I$,

$$\begin{aligned}\big|E\sigma_n^2(u) - \sigma^2(u)\big| &= \sum_{i=1}^n W_{ni}(u)\big([\sigma^2(u)]'(\mu_i - u) + [\sigma^2(\zeta_i)]''(\mu_i - u)^2\big)\\ &\le \sum_{i=1}^n \big|W_{ni}(u)\big|\,\big|[\sigma^2(\zeta_i)]'' - [\sigma^2(u)]''\big|\,(\mu_i - u)^2 + O(b^2)\\ &= O(b^2).\end{aligned}$$

For the mean squared errors, we note that $\operatorname{var}(\sigma_n^2(u)) = O(1/nb)$, which together with the above implies the result.

For the stochastic part, Lemma 6.3 in Müller and Zhao (1995) can be applied by letting

$$\eta(\mu_i) = \delta_i - E\delta_i = \varepsilon_i^2 - \sigma^2(\mu_i).$$

Then $\eta(\mu_i)$, $i = 1, 2, \dots, n$, are independent random variables with $E[\eta(\mu_i)] = 0$ and

$$\varphi(\mu_i) = E[\eta^2(\mu_i)] = E[\varepsilon_i^4] - \big(\sigma^2(\mu_i)\big)^2,$$

for a function $\varphi(\cdot)$. By (M4), the inequalities $\sup_{u\in I}|\varphi(u)| < \infty$ and $E(|\eta(\mu_i)|^s) < c < \infty$ hold. By combining (M2), (M4) and (M5), we find that $\operatorname{var}(\varepsilon_i^2) > 0$, $\varphi(\mu_i) > 0$, φ is continuous and μ_i are in a compact subset of $\mathbb{R}$. This implies that $\inf_{u\in I}|\varphi(u)| > 0$. Combining these observations with (M6), we find that the assumptions of Lemma 6.3 in Müller and Zhao (1995) are satisfied. This leads to

$$\sup_{u\in I}\big|\sigma_n^2(u) - E\sigma_n^2(u)\big| = O_p\left(\left[\frac{\log n}{nb}\right]^{1/2}\right).$$

For any $u \in I$,

$$\big|\sigma_n^2(u) - \sigma^2(u)\big| \le \big|\sigma_n^2(u) - E\sigma_n^2(u)\big| + \big|E\sigma_n^2(u) - \sigma^2(u)\big|,$$

which completes the proof. □

PROOF OF THEOREM 4.1. By decomposition, $\hat{\sigma}_n^2(u)$ in (3.9) can be written as

$$\sum_{i=1}^n \hat{W}_{ni}(u)\,\hat{\varepsilon}_i^2 = \sum_{i=1}^n\big[W_{ni}(u) + \big(\hat{W}_{ni}(u) - W_{ni}(u)\big)\big]\big[\varepsilon_i + (\hat{\varepsilon}_i - \varepsilon_i)\big]^2.$$

Using $\sigma_n^2(u)$ in (3.2),

$$\begin{aligned}\sup_{u\in I}\left|\sum_{i=1}^n \hat{W}_{ni}(u)\,\hat{\varepsilon}_i^2-\sigma_n^2(u)\right| &\le \sup_{u\in I}\sum_{i=1}^n\left|\hat{W}_{ni}(u)-W_{ni}(u)\right|\varepsilon_i^2\\ &\quad+\sup_{u\in I}\sum_{i=1}^n\left|W_{ni}(u)\right|(\hat{\varepsilon}_i-\varepsilon_i)^2\\ &\quad+\sup_{u\in I}\sum_{i=1}^n\left|\hat{W}_{ni}(u)-W_{ni}(u)\right|(\hat{\varepsilon}_i-\varepsilon_i)^2\\ &\quad+2\sup_{u\in I}\sum_{i=1}^n\left|W_{ni}(u)\varepsilon_i(\hat{\varepsilon}_i-\varepsilon_i)\right|\\ &\quad+2\sup_{u\in I}\sum_{i=1}^n\left|(\hat{W}_{ni}(u)-W_{ni}(u))\varepsilon_i(\hat{\varepsilon}_i-\varepsilon_i)\right|\\ &=I+II+III+IV+V.\end{aligned}$$

Now, consider the terms I–V individually. Using Lemmas 8.7, 8.8 and (K3),

$$\begin{aligned}I&=\sup_{u\in I}\sum_{i=1}^n\left|\hat{W}_{ni}(u)-W_{ni}(u)\right|\varepsilon_i^2\\ &\le\left\{\sup_{u\in I}\max_{1\le i\le n}\left|\hat{W}_{ni}(u)-W_{ni}(u)\right|\right\}\sup_{u\in I}\left(\sum_{i=1}^n\varepsilon_i^2 1_{\{|\mu_i-u|\le b\}}\right)\\ &=O_p\left(\frac{1}{b\sqrt{n}}\right).\end{aligned}$$

Writing $\hat{\varepsilon}_i=y_i-\hat{\mu}_i=\varepsilon_i+\mu_i-\hat{\mu}_i$ and using (8.15), that is, $\max_{1\le i\le n}|\hat{\varepsilon}_i-\varepsilon_i|=O_p(1/\sqrt{n})$, along with (8.9),

$$\begin{aligned}II&=\sup_{u\in I}\sum_{i=1}^n\left|W_{ni}(u)\right|(\hat{\varepsilon}_i-\varepsilon_i)^2\\ &\le\max_i\left(|\hat{\varepsilon}_i-\varepsilon_i|^2\right)\sup_{u\in I}\sum_{i=1}^n\left|W_{ni}(u)\right|\\ &=O_p\left(\frac{1}{n}\right).\end{aligned}$$

Similarly, using (8.15) and (8.7),

$$\begin{aligned}III&=\sup_{u\in I}\sum_{i=1}^n\left|\hat{W}_{ni}(u)-W_{ni}(u)\right|(\hat{\varepsilon}_i-\varepsilon_i)^2\\ &\le\sup_{u\in I}\max_i\left|\hat{W}_{ni}(u)-W_{ni}(u)\right|\sup_{u\in I}\sum_{i=1}^n(\hat{\varepsilon}_i-\varepsilon_i)^2\,1_{\{|\hat{\mu}_i-u|\le b\}}\\ &=O_p\left(\frac{1}{nb\sqrt{n}}\right).\end{aligned}$$

Let $W_{ni}^{+}(u) = \max(0,\ W_{ni}(u))$ and $W_{ni}^{-}(u) = \max(0,\ -W_{ni}(u))$. The Cauchy–Schwarz inequality, Lemma 8.8 and the bound for II imply

$$\begin{aligned}
\frac{1}{2}IV &\le \sup_{u\in I}\sum_{i=1}^{n}\big(W_{ni}^{+}(u)+W_{ni}^{-}(u)\big)|\varepsilon_i|\ |\hat{\varepsilon}_i-\varepsilon_i| \\
&\le \sup_{u\in I}\Big(\sum_{i=1}^{n} W_{ni}^{+}(u)\ \varepsilon_i^2\Big)^{1/2}\sup_{u\in I}\Big(\sum_{i=1}^{n} W_{ni}^{+}(u)(\hat{\varepsilon}_i-\varepsilon_i)^2\Big)^{1/2} \\
&\quad+\sup_{u\in I}\Big(\sum_{i=1}^{n} W_{ni}^{-}(u)\ \varepsilon_i^2\Big)^{1/2}\sup_{u\in I}\Big(\sum_{i=1}^{n} W_{ni}^{-}(u)\ (\hat{\varepsilon}_i-\varepsilon_i)^2\Big)^{1/2} \\
&\le \sup_{u\in I}\Big(\max_{1\le i\le n} W_{ni}^{+}(u)\Big)^{1/2}\sup_{u\in I}\Big(\sum_{i=1}^{n}\varepsilon_i^2 1_{\{|\mu_i-u|\le b\}}\Big)^{1/2} O_p\Big(\frac{1}{\sqrt{n}}\Big) \\
&\quad+\sup_{u\in I}\Big(\max_{1\le i\le n} W_{ni}^{-}(u)\Big)^{1/2}\sup_{u\in I}\Big(\sum_{i=1}^{n}\varepsilon_i^2 1_{\{|\mu_i-u|\le b\}}\Big)^{1/2} O_p\Big(\frac{1}{\sqrt{n}}\Big) \\
&= O_p\Big(\frac{1}{\sqrt{n}}\Big).
\end{aligned}$$

Let $w_{ni}^{+}(u) = (\hat{W}_{ni}(u) - W_{ni}(u))^{+}$ and $w_{ni}^{-}(u) = (\hat{W}_{ni}(u) - W_{ni}(u))^{-}$. The Cauchy–Schwarz inequality, Lemma 8.8 and the bound for III imply

$$\begin{aligned}
\frac{1}{2}V &\le \sup_{u\in I}\Big(\sum_{i=1}^{n} w_{ni}^{+}(u)\ \varepsilon_i^2\Big)^{1/2}\sup_{u\in I}\Big(\sum_{i=1}^{n} w_{ni}^{+}(u)(\hat{\varepsilon}_i-\varepsilon_i)^2\Big)^{1/2} \\
&\quad+\sup_{u\in I}\Big(\sum_{i=1}^{n} w_{ni}^{-}(u)\varepsilon_i^2\Big)^{1/2}\sup_{u\in I}\Big(\sum_{i=1}^{n} w_{ni}^{-}(u)(\hat{\varepsilon}_i-\varepsilon_i)^2\Big)^{1/2} \\
&= O_p\Big(\frac{1}{nb}\Big).
\end{aligned}$$

Hence, $I + II + III + IV + V = O_p(1/b\sqrt{n})$. By Lemma 8.6 and Slutsky's theorem, the proof is complete. □

PROOF OF THEOREM 4.2. The proof of the theorem follows the proof of McCullagh (1983) for the asymptotic normality of quasi-likelihood. Let $U(\beta)$ be the quasi-score function, I_β be the "observed" quasi-information matrix of β with correctly specified variance function $\sigma^2(\cdot)$ and i_β be the expected value of I_β. Then, with $Q(\mu, y)$ defined in (2.1),

$$U(\beta)=\frac{\partial Q(\mu, y)}{\partial\beta}=\sum_{i=1}^{n}(y_i-\mu_i)\frac{g'(\eta_i)}{\sigma^2(\mu_i)}x_i=D^T V^{-1}(Y-\mu), \tag{8.18}$$

$$
\begin{aligned}
I_\beta &= -\frac{\partial^2 Q(\mu, y)}{\partial\beta\,\partial\beta^T} = -\frac{\partial U(\beta)}{\partial\eta}\frac{\partial\eta}{\partial\beta^T}\\
&= \sum_{i=1}^n \frac{g'(\eta_i)^2}{\sigma^2(\mu_i)} x_i x_i^T \\
&\quad + \sum_{i=1}^n (y_i - \mu_i)\left\{\frac{g'(\eta_i)^2[\sigma^2(\mu_i)]'}{[\sigma^2(\mu_i)]^2} - \frac{g''(\eta_i)}{\sigma^2(\mu_i)}\right\} x_i x_i^T,
\end{aligned}
\tag{8.19}
$$

$$
i_\beta = \sum_{i=1}^n \frac{g'(\eta_i)^2}{\sigma^2(\mu_i)} x_i x_i^T = D^T V^{-1} D,
\tag{8.20}
$$

where $x_i = (1, x_{i1}, x_{i2}, \dots, x_{i(p-1)})^T$, and D, V, and Σ are defined in (M7). A QL estimator $\hat{\beta}$ of β is obtained by solving $U(\hat{\beta}) = 0$. By a first-order Taylor series expansion in $(\hat{\beta} - \beta)$, the equation can be written as

$$
U(\beta) - I_{\bar{\beta}}(\hat{\beta} - \beta) = 0.
\tag{8.21}
$$

Here,

$$
I_{\bar{\beta}} = \sum_{i=1}^n \frac{g'(\bar{\eta}_i)^2}{\sigma^2(\bar{\mu}_i)} x_i x_i^T + \sum_{i=1}^n (y_i - \bar{\mu}_i)\left\{\frac{g'(\bar{\eta}_i)^2[\sigma^2(\bar{\mu}_i)]'}{[\sigma^2(\bar{\mu}_i)]^2} - \frac{g''(\bar{\eta}_i)}{\sigma^2(\bar{\mu}_i)}\right\} x_i x_i^T,
\tag{8.22}
$$

and $U(\beta)$ is as in (8.18) with $\bar{\eta}_i = x_i^T\bar{\beta}$, $\bar{\mu}_i = g(\bar{\eta}_i)$, and $I_{\bar{\beta}}$, the "observed" nonparametric quasi-information matrix, is evaluated at a vector $\bar{\beta}$ lying on the line segment joining β and $\hat{\beta}$. Under (M1)–(M5) and (M7), (8.21) leads to the result of (9) in McCullagh (1983), implying

$$
\sqrt{n}(\hat{\beta} - \beta) \to_D N_p(\mathbf{0}, \Sigma^{-1})
\tag{8.23}
$$

where $\hat{\beta}$ is a QLE of β and Σ is defined in (M7).

A NQLE $\hat{\beta}^*$ is obtained by solving the nonparametric quasi-score equation [see (2.3)] $U^*(\hat{\beta}^*) = 0$. By a first-order Taylor series expansion in $(\hat{\beta}^* - \beta)$, the estimating equation $U^*(\hat{\beta}^*) = 0$ as in (2.3) can be written as

$$
U^*(\beta) - I^*_{\bar{\beta}}(\hat{\beta}^* - \beta) = 0.
\tag{8.24}
$$

Here

$$
I^*_{\bar{\beta}} = \sum_{i=1}^n \frac{g'(\bar{\eta}_i)^2}{\hat{\sigma}_n^2(\bar{\mu}_i)} x_i x_i^T + \sum_{i=1}^n (y_i - \bar{\mu}_i)\left\{\frac{g'(\bar{\eta}_i)^2[\hat{\sigma}_n^2(\bar{\mu}_i)]'}{[\hat{\sigma}_n^2(\bar{\mu}_i)]^2} - \frac{g''(\bar{\eta}_i)}{\hat{\sigma}_n^2(\bar{\mu}_i)}\right\} x_i x_i^T,
\tag{8.25}
$$

where $I^*_{\bar{\beta}}$ is the "observed" nonparametric quasi-information matrix evaluated at $\bar{\beta}$ lying on the line segment joining β and $\hat{\beta}^*$. To infer the asymptotic limiting distribution of the solution $\hat{\beta}^*$ of (8.24), we aim to show that it is the same as that of the solution $\hat{\beta}$ of (8.21). This follows if $\sqrt{n}(\hat{\beta}^* - \hat{\beta}) = o_p(1)$. For this it is sufficient to prove $U^*(\beta) - U(\beta) = o_p(\sqrt{n})$ and $I^*_{\bar{\beta}} - I_{\bar{\beta}} = o_p(n)$.

Since the estimating equation is highly nonlinear, an iterative procedure to obtain the solution $\hat{\beta}^*$ is necessary. As an initial step, initial variance function estimates are assigned to obtain the one-step estimated vector of the regression coefficients. Although the initial variance function estimates may be misspecified, a $\sqrt{n}$-consistent estimator $\hat{\beta}^*$ of β is still obtained, but with loss of efficiency. Details regarding the estimation of misspecified generalized linear models can be found in Fahrmeir (1990). Therefore, the vector of the mean value $\bar{\beta}$ in I^*_{β} as in (8.25) satisfies $\|\bar{\beta} - \beta\| = O_p(1/\sqrt{n})$, and it follows by Theorem 4.1 that $\max_{1\le i\le n} |\hat{\sigma}^2_n(\bar{\mu}_i) - \sigma^2(\bar{\mu}_i)| = O_p(\alpha_n)$ where $\alpha_n = b^2 + ((\log n)/nb)^{1/2} + 1/\sqrt{n}b$, $\hat{\sigma}^2_n(u) = \hat{W}_{ni}(u)(y_i - \bar{\mu}_i)^2$, with $\hat{W}_{ni}(u) = W_{ni}(u; \bar{\mu}_1, \bar{\mu}_2, \dots, \bar{\mu}_n)$, $\bar{\mu}_i = g(\bar{\eta}_i)$, and $\bar{\eta}_i = x_i^T\bar{\beta}$.

To simplify the notations, let

$$A_i = \frac{1}{\hat{\sigma}^2_n(\bar{\mu}_i)} - \frac{1}{\sigma^2(\bar{\mu}_i)} \quad \text{and} \quad B_i = \frac{[\hat{\sigma}^2_n(\bar{\mu}_i)]'}{[\hat{\sigma}^2_n(\bar{\mu}_i)]^2} - \frac{[\sigma^2(\bar{\mu}_i)]'}{[\sigma^2(\bar{\mu}_i)]^2}.$$

With the lower bound $\hat{\sigma}^2_n(\cdot) \ge \zeta_n$, we have $\max_{1\le i\le n} |A_i| = O_p(\alpha_n/\zeta_n)$. Furthermore, by extending Theorem 4.1, we obtain $\max_{1\le i\le n} |B_i| = O_p(\alpha_n/\zeta_n b)$. Under the boundedness conditions of (M3) and (M5), and by (8.22) and (8.25), we then have

$$\begin{aligned}\frac{1}{n}\big(I^*_{\bar{\beta}} - I_{\bar{\beta}}\big) &= \frac{1}{n}\sum_{i=1}^{n} A_i\ g'(\bar{\eta}_i)^2 x_i x_i^T \\ &\quad + \frac{1}{n}\sum_{i=1}^{n}(y_i - \bar{\mu}_i)\big\{B_i\ g'(\bar{\eta}_i)^2 - A_i\ g''(\bar{\eta}_i)\big\}x_i x_i^T \\ &\le \frac{1}{n}\sum_{i=1}^{n}\max_{1\le i\le n}|A_i| + \frac{1}{n}\sum_{i=1}^{n}\big|y_i - \bar{\mu}_i\big|\Big(\max_{1\le i\le n}|B_i| + \max_{1\le i\le n}|A_i|\Big) \\ &= o_p(1),\end{aligned} \tag{8.26}$$

using (K4).

Similarly, by (M3) and (M5), we find by a Taylor expansion of $1/\sigma^2(\cdot)$ and some algebra that

$$\begin{aligned}\frac{1}{\sqrt{n}}\big(U^*(\beta) - U(\beta)\big) &= \frac{1}{\sqrt{n}}\sum_{i=1}^{n}(y_i - \mu_i)\bigg(\frac{g'(\eta_i)}{\sigma^2_n(\mu_i)} - \frac{g'(\eta_i)}{\sigma^2(\mu_i)}\bigg)x_i \\ &= O_p\bigg(\frac{1}{\zeta_n\sqrt{n}}\sum_{i=1}^{n}(y_i - \mu_i)\big(\sigma^2(\mu_i) - \sigma^2_n(\mu_i)\big)\bigg).\end{aligned}$$

The desired result

$$\frac{1}{\zeta_n\sqrt{n}}\sum_{i=1}^{n}\varepsilon_i\big(\sigma^2(\mu_i) - \sigma^2_n(\mu_i)\big) = o_p(1), \tag{8.27}$$

where $\varepsilon_i = y_i - \mu_i$, will follow if we show that

$$\begin{aligned}
&\frac{1}{n}E\left\{\left[\sum_{i=1}^{n}\varepsilon_i\Big(\sigma^2(\mu_i)-\sigma_n^2(\mu_i)\Big)\right]^2\right\}\\
&\quad=\frac{1}{n}\sum_{i=1}^{n}\left\{E\left[\varepsilon_i^2\Big(\sigma^2(\mu_i)-\sigma_n^2(\mu_i)\Big)^2\right]\right\}\\
&\qquad+\frac{2}{n}\sum_{1\le j<k\le n}\left\{E\left[\varepsilon_j\Big(\sigma^2(\mu_j)-\sigma_n^2(\mu_j)\Big)\varepsilon_k\Big(\sigma^2(\mu_k)-\sigma_n^2(\mu_k)\Big)\right]\right\}
\end{aligned}\tag{8.28}$$

is $o_p(\zeta_n^2)$. For this, let $\sigma_n^2(u)^{(-i)} = \sigma_n^2(u) - W_{ni}(u)\varepsilon_i^2$ be the leave-one-point-out variance function estimator. By a straightforward extension of Lemma 4.1, we find for all $1 \le i \le n$,

$$E\big[\sigma_n^2(u)^{(-i)} - \sigma^2(u)\big] = O\bigg(\frac{1}{nb} + b^2\bigg)$$

and

$$E\big[(\sigma_n^2(u)^{(-i)} - \sigma^2(u))^2\big] = O\bigg(\frac{1}{nb} + b^4\bigg).$$

By substituting $\sigma_n^2(\mu_i)$ with $(\sigma_n^2(\mu_i)^{(-i)} + W_{ni}(\mu_i)\varepsilon_i^2)$, we have, for all $1 \le i \le n$,

$$\begin{aligned}
&E\left[\varepsilon_i^2\ \Big(\sigma^2(\mu_i)-\sigma_n^2(\mu_i)\Big)^2\right]\\
&\quad= E\varepsilon_i^2\ E\big[(\sigma_n^2(u)^{(-i)} - \sigma^2(u))^2\big] + W_{ni}(\mu_i)^2\ E\varepsilon_i^6\\
&\qquad+ 2W_{ni}(\mu_i)\ E\varepsilon_i^4\ E\Big[\sigma_n^2(u)^{(-i)} - \sigma^2(u)\Big]\\
&\quad= O\bigg(\frac{1}{nb}+b^4\bigg) + O\bigg(\frac{1}{n^2b^2}\bigg) + O\bigg(\frac{1}{nb}\bigg)O\bigg(\frac{1}{nb}+b^2\bigg).
\end{aligned}$$

Let $\sigma_n^2(u)^{(-j,-k)} = \sigma_n^2(u) - W_{nj}(u)\varepsilon_j^2 - W_{nk}(u)\varepsilon_k^2$, $j < k$, be the leave-two-points-out variance function estimator. By substituting $\sigma_n^2(\mu_i)$ with $(\sigma_n^2(\mu_i)^{(-j,-k)} + W_{nj}(\mu_i)\varepsilon_j^2 + W_{nk}(\mu_i)\varepsilon_k^2)$, for all $1 \le i \le n$ and $1 \le j < k \le n$, we obtain

$$\begin{aligned}
&E\big[\varepsilon_j(\sigma^2(\mu_j)-\sigma_n^2(\mu_j))\varepsilon_k(\sigma^2(\mu_k)-\sigma_n^2(\mu_k))\big]\\
&\quad= W_{nj}(\mu_i)\ W_{nk}(\mu_k)\ E\varepsilon_j^3\ E\varepsilon_k^3\\
&\quad= O\left(\frac{1}{n^2b^2}\right).
\end{aligned}$$

This implies with (K4) that the terms in (8.28) are indeed of order $o_p(\zeta_n^2)$, and (8.27) follows. We infer

$$\frac{1}{\sqrt{n}}\big(U^*(\beta) - U(\beta)\big) = o_p(1). \tag{8.29}$$

By (8.29) and (8.26), we get from (8.24),

$$U^*(\beta) - I^*_{\tilde{\beta}}(\hat{\beta}^* - \beta) = U(\beta) - I_{\tilde{\beta}}(\hat{\beta}^* - \beta) + o_p(\sqrt{n}).$$

Comparing to (8.21) and (12) and (13) in McCullagh (1983), we finally arrive at

$$\sqrt{n}(\hat{\beta}^* - \beta) = \sqrt{n}(\hat{\beta} - \beta) + o_p(1). \qquad \square$$

Acknowledgment. We gratefully acknowledge the constructive remarks of a referee which led to several improvements in the paper.

REFERENCES

ATKINSON, A. C. (1987). *Plots, Transformations, and Regression.* Clarendon Press, Oxford.

CARROLL, R. J. and RUPPERT, D. (1982). Robust estimation in heteroscedastic linear models. *Ann. Statist.* **10** 429–441.

CHIOU, J.-M. and MÜLLER, H.-G. (1998). Quasi-likelihood regression with unknown link and variance functions. *J. Amer. Statist. Assoc.* **93** 1376–1387.

DAVIDIAN, M. and CARROLL, R. J. (1988). A note on extended quasi-likelihood. *J. Roy. Statist. Soc. Ser. B* **50** 74–82.

FAHRMEIR, L. (1990). Maximum likelihood estimation in misspecified generalized linear models. *Statistics* **21** 487–502.

FAN, J. and GIJBELS, I. (1996) *Local Polynomial Modeling and Its Applications*. Chapman and Hall, London.

FAN, J. and TRUONG, Y. K. (1993) Nonparametric regression with errors-in-variables. *Ann. Statist.* **21** 1900–1925.

HALL, P. and CARROLL, R. J. (1989). Variance function estimation: The effect of estimating the mean. *J. Roy. Statist. Soc. Ser. B* **51** 3–14.

HALL, P., KAY, J. W. and TITTERINGTON, D. M. (1990). Asymptotically optimal difference-based estimation of variance in nonparametric regression. *Biometrika* **77** 521–528.

MCCULLAGH, P. (1983). Quasi-likelihood functions. *Ann. Statist.* **11** 59–67.

MCCULLAGH, P. and NELDER, J. A. (1989) *Generalized Linear Models.* Chapman and Hall, London.

MÜLLER, H.-G. (1988) Nonparametric regression analysis of longitudinal data. *Lecture Notes in Statist.* **46**. Springer, New York.

MÜLLER, H.-G. and STADTMÜLLER, U. (1987). Estimation of heteroscedasticity in regression analysis. *Ann. Statist.* **15** 610–625.

MÜLLER, H.-G. and STADTMÜLLER, U. (1993). On variance function estimation with quadratic forms. *J. Statist. Plann. Inference* **35** 213–231.

MÜLLER, H.-G. and ZHAO, P.-L. (1995). On a semiparametric variance function model and a test for heteroscedasticity. *Ann. Statist.* **23** 946–967.

NELDER, J. A. and LEE, Y. (1992). Likelihood, quasi-likelihood and pseudo-likelihood: Some comparisons. *J. Roy. Statist. Soc. Ser. B* **54** 273–284.

NELDER, J. A. and PREGIBON, P. (1987). An extended quasi-likelihood function. *Biometrika* **74** 221–232.

NELDER, J. A. and WEDDERBURN, R. W. M. (1972) Generalized linear models. *J. Roy. Statist. Soc. Sec. A* **135** 370–384.

RUPPERT, D., WAND, M. P., HOLST, U. and HÖSSJER, O. (1997). Local polynomial variance function estimation. *Technometrics* **39** 262–273.

RYAN, T., JOINER, B. and RYAN, B. (1985). *Minitab Student Handbook*, 2nd ed. Duxbury, Belmont, CA.

SILVERMAN, B. W. (1985). Some aspects of the spline smoothing approach to non-parametric regression curve fitting (with discussion). *J. Roy. Statist. Soc. Ser. B* **47** 1–52.

THOMPSON, A. M., BROWN, J. C., KAY, J. W. and TITTERINGTON, D. M. (1991). A study of methods of choosing the smoothing parameter in image restoration by regularization. *IEEE Trans. Pattern Recognition Machine Intell.* **13** 326–338.

TITTERINGTON, D. M. (1985). Common structure of smoothing techniques in statistics. *Internat. Statist. Rev.* **53** 141–170.

WEDDERBURN, R. W. M. (1974). Quasi-likelihood functions, generalized linear models and Gauss–Newton method. *Biometrika* **61** 439–447.

WEISBERG, S. and WELSH, A. H. (1994). Adapting for the missing link. *Ann. Statist.* **22** 1674–1700.

DEPARTMENT OF MATHEMATICS
NATIONAL CHUNG CHENG UNIVERSITY
MINGHSIUNG
CHIAYI 621
TAIWAN
E-MAIL: jmchiou@math.ccu.edu.tw

DIVISION OF STATISTICS
1 SHIELDS AVENUE
UNIVERSITY OF CALIFORNIA
DAVIS, CALIFORNIA 95616
E-MAIL: mueller@wald.ucdavis.edu