Variance reduction in multiparameter likelihood models

Ming-Yen Cheng and Liang Peng

May 2006

Abstract

Local likelihood modeling is a unified and effective approach to establishing the dependence of a response variable, which can be of various types, on independent variables. Therefore the methods have become popular in a wide range of applications. There is an increasing interest in employing multiparameter local likelihood models to investigate trends of sample extremes in environmental statistics. When sample maxima are modeled by a generalized extreme value distribution, the sample size is small in general and local likelihood estimation exhibits a large variation. In this paper, variance reduction techniques are employed to improve efficiency of the inference. A simulation study and an application to annual maximum temperatures show that our methods are very effective in finite samples.

Key Words. Bootstrap, extreme value distribution, generalized linear models, local likelihood, local linear MLE, logistic regression.

Ming-Yen Cheng is Professor, Department of Mathematics, National Taiwan University, Taipei 106, Taiwan (email: cheng@math.ntu.edu.tw); and Liang Peng is Associate Professor, School of Mathematics, Georgia Institute of Technology, Atlanta GA 30332-0160, USA (email: peng@math.gatech.edu). This work was partially supported by NSF grant DMS0403443 and a Humboldt research fellowship. The comments from an editor, an associate editor and two reviewers are greatly appreciated.

1. Introduction

Suppose that $(X_1, Y_1), \dots, (X_n, Y_n)$ are independent bivariate observations from the distribution of (X, Y) . Consider a multiparameter likelihood model

$$f(y; \theta_1(x), \dots, \theta_d(x)) \quad (1)$$

for the conditional density of the response Y given that the covariate $X = x$, where the form of the probability density function f is known and the unknown parameters $\theta(x) = (\theta_1(x), \dots, \theta_d(x))^T$ depend on $X = x$. Under model (1), dependence of Y on X is specified by a parametric law, with probability density function f , in which the parameter vector θ being an unknown d -dimensional function of X .

Statistical inference for the underlying population (X, Y) based on the observed data $(X_1, Y_1), \dots, (X_n, Y_n)$ heavily relies on nonparametric estimation of the curves $\theta_1(x), \dots, \theta_d(x)$. An efficient approach is the local linear maximum likelihood estimation: locally around every x the curves $\theta_1(\cdot), \dots, \theta_d(\cdot)$ are modeled as linear functions and then estimated by maximizing a kernel weighted likelihood function. Specifically, define the local linear log-likelihood at $\theta(x)$ as

$$L_n(\theta^*(x)) = \sum_{i=1}^n K_h(X_i - x) \log f(Y_i; \theta(x, X_i)),$$

where $K_h(x) = K(x/h)/h$, K is a kernel, $h > 0$, $h = h(n) \rightarrow 0$ as $n \rightarrow \infty$, is a bandwidth, $\theta^*(x) = (\theta_1(x), \theta'_1(x), \dots, \theta_d(x), \theta'_d(x))^T$, $\theta(x, u) = \theta(x) + (\theta'_1(x)(u - x), \dots, \theta'_d(x)(u - x))^T$. Then local linear maximum likelihood estimator $\hat{\theta}^*(x) = (\hat{\theta}_1(x), \hat{\theta}'_1(x), \dots, \hat{\theta}_d(x), \hat{\theta}'_d(x))^T$ of $\theta^*(x)$ maximizes $L_n(\theta^*(x))$, i.e.,

$$\hat{\theta}^*(x) = \arg \max_{\theta^*(x)} L_n(\theta^*(x)).$$

Note that in model (1) what is essential is $\theta(x) = (\theta_1(x), \dots, \theta_d(x))^T$ and the derivatives $\alpha(x) = (\theta'_1(x), \dots, \theta'_d(x))^T$ are irrelevant. Hence, the local linear maximum

likelihood estimator for $\theta(x)$ is defined as

$$\hat{\theta}(x) = (\hat{\theta}_1(x), \hat{\theta}_2(x), \dots, \hat{\theta}_d(x))^T. \quad (2)$$

We refer to Aerts and Claeskens (1997) and Fan, et al. (1998) for the asymptotic properties of $\hat{\theta}^*(x)$ and $\hat{\theta}(x)$ and the choice of bandwidth. In addition, see Claeskens and Van Keilegom (2003) for a study on confidence bands of $\theta(x)$.

To provide a motivating example for studying variance reduction in construction of multiparameter local likelihood models, we mention modeling of extremes and exceedances. Recently, there is an increasing interest in applying likelihood models to investigate the trend in sample extremes; see Davison and Ramesh (2000) for fitting a generalized extreme value distribution to sample maxima, Hall and Tajvidi (2000) for fitting a generalized extreme value distribution and a generalized Pareto distribution to data, and Beirlant and Goegebeur (2004) for fitting a generalized Pareto distribution to exceedances by taking the unknown high threshold into account. Some applications of fitting an extreme value distribution locally to environmental data can be found in Ramesh and Davison (2002) and Chavez-Demoulin and Davison (2005). When we model sample maxima by a generalized extreme value distribution, the sample size is not large in general. Then the estimation procedure can become much more reliable and efficient provided that variance reduction techniques are implemented.

Kogure (1998) studied general order polynomial interpolation of kernel density estimation and showed that the asymptotic integrated variance becomes smaller. Cheng, et al. (2005) introduced variance reduction techniques for nonparametric regression which reduce the pointwise asymptotic variance uniformly. In this paper we adopt the approach of Cheng, et al. (2005) since it is more effective. Theoretical study shows that our method reduces asymptotic variances of the d parameter estimators by a common and known constant factor. Interestingly, this variance reduction in estimating d parameters is simultaneously achieved by applying the technique once

all together. These results are nontrivial given those of Cheng, et al. (2005): multi-parameter local likelihood model specifies the conditional distribution of the response variable given the covariates while nonparametric regression considers the conditional mean, and asymptotic behaviors of nonparametric kernel regression are different from those of local likelihood estimation.

Here we discuss briefly the scope of local likelihood models outside extremes and exceedances. Local likelihood methods effectively model dependence of various kinds of response variables on covariates in a unified framework. If $f(y; \theta_1(x)) = \exp[C + \{y - \theta_1(x)\}^2/\sigma]$, where C and σ are given constants, then the local linear MLE reduces to the local linear regression estimator, see e.g. Loader (1999). Yu and Jones (2004) adapted an analogous local Normal likelihood model for estimation of the conditional variance function in nonparametric regression. Tibshirani and Hastie (1987) suggested (2) when $d = 1$ and applied to local logistic regression and local partial likelihood estimation of Cox's proportional hazard model. Staniswalis (1989) considered local constant MLE and allowed X_i to be multivariate. Another special case of local likelihood modeling is, in generalized linear models, when the conditional distribution of Y given X belongs to a one-parameter exponential family and the parameter depends on X . In this regard, Loader (1999) discussed various examples including local Poisson regression and local Gamma model for survival analysis. Fan, et al. (1995) extended the idea to local quasi-likelihood estimation. The literature further includes Irizarry (2001), Eguchi, et al. (2003) and Signorini and Jones (2004).

We organize this paper as the following. In Section 2, we demonstrate a way to incorporate the variance reduction techniques of Cheng, et al. (2005) to improve the local linear maximum likelihood estimator (2). In addition, the main theoretical results and discussions on further variants are suggested. A simulation study and a real application on annual maximum temperatures are presented in Section 3. All

proofs of the theoretical results are given in the appendix.

2. Methodology and Main Results

The idea of our variance reduction strategy is as follows. For each point of estimation, construct a linear combination of local linear maximum likelihood estimates at three points around the point such that the asymptotic bias remains unchanged. Specifically, for any given point x , let $\{\beta_{x,0}, \beta_{x,1}, \beta_{x,2}\}$ be an equally spaced grid of points with bin width $\delta h = \beta_{x,1} - \beta_{x,0} = \beta_{x,2} - \beta_{x,1}$ such that $x = \beta_{x,1} + r\delta h$, where $r \in (-1, 1) \setminus \{0\}$ and $\delta > 0$ are given constants. Then, like Cheng, et al. (2005), a variance reduction estimator for $\theta(x)$ is defined as

$$\tilde{\theta}(x) = \frac{r(r-1)}{2} \hat{\theta}(\beta_{x,0}) + (1-r^2) \hat{\theta}(\beta_{x,1}) + \frac{r(r+1)}{2} \hat{\theta}(\beta_{x,2}), \quad (3)$$

where $\hat{\theta}(x) = (\hat{\theta}_1(x), \dots, \hat{\theta}_d(x))^T$ is the local linear maximum likelihood estimate given in (2). Suppose that $Supp(X)$ is bounded, $Supp(X) = [0, 1]$ say, since $x - (1-r)\delta h = \beta_{x,0} < x < \beta_{x,2} = x + (1+r)\delta h$, the grid points $\beta_{x,0}$ and $\beta_{x,2}$ would be outside $Supp(X)$ if x is close to the endpoints. Therefore we take

$$\delta(x) = \min \left\{ \delta, x/((1+r)h), (1-x)/((1-r)h) \right\}$$

such that $\{\beta_{x,0}, \beta_{x,1}, \beta_{x,2}\} \in Supp(X) = [0, 1]$ all the time.

Next we compare asymptotic distributions of our variance reduction estimator $\tilde{\theta}(x)$ and the local linear maximum likelihood estimator $\hat{\theta}(x)$. For simplicity, we consider the case $d = 2$. Generalization of the results to general d values is straightforward. Define the local Fisher information matrix of $\theta(x) = (\theta_1(x), \theta_2(x))^T$ as

$$I(\theta_1(x), \theta_2(x)) = \begin{pmatrix} I_{11}(\theta_1(x), \theta_2(x)) & I_{12}(\theta_1(x), \theta_2(x)) \\ I_{21}(\theta_1(x), \theta_2(x)) & I_{22}(\theta_1(x), \theta_2(x)) \end{pmatrix}$$

where

$$\begin{aligned} I_{st}(\theta_1(x), \theta_2(x)) &= \mathbb{E}_x \left\{ -\frac{\partial^2}{\partial \theta_s \partial \theta_t} \log f(Y; \theta_1(x), \theta_2(x)) \right\} \\ &= \mathbb{E}_x \left\{ \frac{\partial}{\partial \theta_s} \log f(Y; \theta_1(x), \theta_2(x)) \frac{\partial}{\partial \theta_t} \log f(Y; \theta_1(x), \theta_2(x)) \right\}, \end{aligned}$$

and $\mathbb{E}_x$ denotes the expectation conditional on $X = x$. Let $f_X(x)$ denote the marginal probability density function of X . Define $\nu_{i,j} = \int z^i K^j(z) dz$, $C(s, t) = \int K(u - st)K(u + st) du$ and $C(s) = \frac{3}{2}C(0, s) - 2C(\frac{1}{2}, s) + \frac{1}{2}C(1, s)$. The following theorem states the asymptotic normality of our variance reduction estimator $\tilde{\theta}(x)$.

Theorem 1. *Under the same regularity conditions as given by Aerts and Claeskens (1997), for interior point x we have, as $n \rightarrow \infty$,*

$$\sqrt{nh} \{ \tilde{\theta}(x) - \theta(x) - \frac{1}{2} h^2 \nu_{2,1} \theta''(x) \} \xrightarrow{d} Z_1, \quad (4)$$

where $\theta''(x) = (\theta_1''(x), \theta_2''(x))^T$ and Z_1 is a d -dimensional normal random vector with zero mean and covariance matrix $\{ \nu_{0,2} - r^2(1 - r^2)C(\delta) \} f_X(x)^{-1} I(\theta_1(x), \theta_2(x))^{-1}$.

It follows from Aerts and Claeskens (1997) that, as $n \rightarrow \infty$,

$$\sqrt{nh} \{ \hat{\theta}(x) - \theta(x) - \frac{1}{2} h^2 \nu_{2,1} \theta''(x) \} \xrightarrow{d} Z_2 \quad (5)$$

where Z_2 is a d -dimensional normal random vector with zero mean and covariance matrix $\nu_{0,2} f_X(x)^{-1} I(\theta_1(x), \theta_2(x))^{-1}$

Remark 1. When x is a boundary point; i.e. x is close to the endpoints of $\text{Supp}(X)$, $\hat{\theta}(x)$ and $\tilde{\theta}(x)$ each still has an asymptotic normal distribution and only the constant factors in the asymptotic bias vector and covariance matrix change. Typically, the asymptotic variances are inflated due to reduced number of data points there.

Define the asymptotic mean squared error (AMSE) of an estimator of $\theta(x)$ as sum of trace of the covariance matrix and squared norm of the asymptotic bias vector in

its asymptotic Normal distribution. Then, from (4) and (5), the asymptotic mean squared errors of $\tilde{\theta}(x)$ and $\hat{\theta}(x)$ are respectively

$$\begin{aligned} \text{AMSE}\{\tilde{\theta}(x)\} &= \{nhf_X(x)\}^{-1} \text{tr}\{I(\theta_1(x), \theta_2(x))^{-1}\} \{\nu_{0,2} - r^2(1 - r^2)C(\delta)\} \\ &\quad + \frac{1}{4}h^4\nu_{2,1}^2\{(\theta_1''(x))^2 + (\theta_2''(x))^2\}, \end{aligned} \quad (6)$$

$$\begin{aligned} \text{AMSE}\{\hat{\theta}(x)\} &= \{nhf_X(x)\}^{-1} \text{tr}\{I(\theta_1(x), \theta_2(x))^{-1}\} \nu_{0,2} \\ &\quad + \frac{1}{4}h^4\nu_{2,1}^2\{(\theta_1''(x))^2 + (\theta_2''(x))^2\}. \end{aligned} \quad (7)$$

Comparing (6) and (7), the asymptotic mean squared error of $\tilde{\theta}(x)$ differs from that of $\hat{\theta}(x)$ by the term $-r^2(1 - r^2)C(\delta)\{nhf_X(x)\}^{-1} \text{tr}\{I(\theta_1(x), \theta_2(x))^{-1}\}$. Note that $0 < r^2(1 - r^2) \leq 1/4$ for any $r \in (-1, 1) \setminus \{0\}$ and it attains the maximum at $r = \pm 2^{-1/2}$. Moreover, for any symmetric kernel K , $0 \leq C(\delta) \leq 3\nu_{0,2}/2$ for all $\delta > 0$ and $C(\delta)$ is increasing in δ if K is in addition unimodal and concave; see Cheng, et al. (2005). Hence, the variance reduction estimator is better than the local linear maximum likelihood estimator in terms of asymptotic mean squared errors.

Remark 2. Comparing (4) and (5), our variance reduction method simultaneously reduces asymptotic variances in estimating all the parameters $\theta_1(x), \dots, \theta_d(x)$ no matter what d , number of parameters in the local likelihood model, is. This property holds for all our estimators, $\tilde{\theta}^{(j)}(x)$, $j = 1, 2, 3$, which are introduced in (8), (9) and (16) below and are constructed based on $\tilde{\theta}(x)$.

Note that r in the definition of $\tilde{\theta}(x)$ is an arbitrary constant in $(-1, 1) \setminus \{0\}$. As discussed earlier, choosing $r = \pm 2^{-1/2}$ we achieve the most variance reduction regardless what h , δ and K are and the resultant estimators are

$$\tilde{\theta}^{(1)}(x) = \frac{1 - 2^{1/2}}{4} \hat{\theta}(x - (1 + 2^{-1/2})\delta h) + \frac{1}{2} \hat{\theta}(x - 2^{-1/2}\delta h) + \frac{1 + 2^{1/2}}{4} \hat{\theta}(x - (2^{-1/2} - 1)\delta h) \quad (8)$$

and

$$\tilde{\theta}^{(2)}(x) = \frac{1 + 2^{1/2}}{4} \hat{\theta}(x + (2^{-1/2} - 1)\delta h) + \frac{1}{2} \hat{\theta}(x + 2^{-1/2}\delta h) + \frac{1 - 2^{1/2}}{4} \hat{\theta}(x + (2^{-1/2} + 1)\delta h). \quad (9)$$

The next theorem states the asymptotic mean squared error of the above variance reduction estimators.

Theorem 2. *Under the same regularity conditions as given by Aerts and Claeskens (1997), for interior point x we have, as $n \rightarrow \infty$,*

$$\begin{aligned} \text{AMSE}\{\tilde{\theta}^{(j)}(x)\} &= \{nhf_X(x)\}^{-1} \text{tr}\{I(\theta_1(x), \theta_2(x))^{-1}\} \left\{ \nu_{0,2} - \frac{C(\delta)}{4} \right\} \\ &\quad + \frac{1}{4} h^4 \nu_{2,1}^2 \{(\theta_1''(x))^2 + (\theta_2''(x))^2\}, \end{aligned} \quad (10)$$

$j = 1, 2$.

Remark 3. It follows from (7) that the optimal bandwidth minimizing $\text{AMSE}\{\hat{\theta}(x)\}$ is

$$h_0 = \left\{ \frac{\nu_{0,2}}{f_X(x)\nu_{2,1}^2} \right\}^{1/5} \left\{ \frac{\text{tr}\{I(\theta_1(x), \theta_2(x))^{-1}\}}{(\theta_1''(x))^2 + (\theta_2''(x))^2} \right\}^{1/5} n^{-1/5}. \quad (11)$$

Similarly, for $j = 1, 2$, the optimal bandwidth minimizing $\text{AMSE}\{\tilde{\theta}^{(j)}(x)\}$ given in (10) is

$$h_j = \{\nu_{0,2} - C(\delta)/4\}^{1/5} \nu_{0,2}^{-1/5} h_0. \quad (12)$$

Remark 4. The bandwidth h_0 given in (11) yields the optimal AMSE of $\hat{\theta}(x)$:

$$\text{AMSE}_0(x) = \left\{ \frac{\nu_{0,2} \text{tr}\{I(\theta_1(x), \theta_2(x))^{-1}\}}{f_X(x)} \right\}^{4/5} \left\{ \nu_{2,1}^2 [(\theta_1''(x))^2 + (\theta_2''(x))^2] \right\}^{1/5} n^{-4/5}. \quad (13)$$

For $j = 1, 2$, the optimal bandwidth given in (12) yields the optimal AMSE of $\tilde{\theta}^{(j)}(x)$:

$$\text{AMSE}_j(x) = \{\nu_{0,2} - C(\delta)/4\}^{4/5} \nu_{0,2}^{-4/5} \text{AMSE}_0(x), \quad (14)$$

and hence the asymptotic relative efficiency of $\tilde{\theta}^{(j)}(x)$ compared to $\hat{\theta}(x)$ is

$$\text{eff}\{\tilde{\theta}^{(j)}(x), \hat{\theta}(x)\} = \{\nu_{0,2} - C(\delta)/4\}^{-4/5} \nu_{0,2}^{4/5} \geq 1. \quad (15)$$

While theoretical results imply that more variance reduction is achieved by implementing larger δ values, that may introduce large finite sample bias effects. We suggest to take $\delta = 1$ for general purposes. Slightly larger values of δ , e.g. $\delta = 1.2$, may be useful in applications when the second derivatives of the curves $\theta_j(x)$, $j = 1, \dots, d$ are small. A simple way to judge is to examine departures of the curve estimates $\tilde{\theta}^{(1)}(\cdot)$ and $\tilde{\theta}^{(2)}(\cdot)$ from $\hat{\theta}(\cdot)$ when using different δ choices.

Either of the variance reduction estimator $\tilde{\theta}^{(1)}(x)$ or $\tilde{\theta}^{(2)}(x)$ uses more information from data points on one side of x than those on the other side; see (8) and (9). One way to cancel out these finite sample biases is to take the average of the two estimators

$$\tilde{\theta}^{(3)}(x) = \frac{1}{2} \{ \tilde{\theta}^{(1)}(x) + \tilde{\theta}^{(2)}(x) \}. \quad (16)$$

When $\text{Supp}(X) = [0, 1]$, to keep the points $\{\beta_{x,0}, \beta_{x,1}, \beta_{x,2}\}$ with both $r = 2^{-1/2}$ and $r = -2^{-1/2}$ all within the data range $[0, 1]$ we let

$$\delta(x) = \min \{ \delta, x / ((1 + 2^{-1/2})h), (1 - x) / ((1 + 2^{-1/2})h) \}$$

for a given positive constant δ , $\delta = 1$ say.

Theorem 3. *Under the same regularity conditions as of Aerts and Claeskens (1997), for interior point x we have, as $n \rightarrow \infty$,*

$$\sqrt{nh} \{ \tilde{\theta}^{(3)}(x) - \theta(x) - \frac{1}{2} h^2 \nu_{2,1} \theta''(x) \} \xrightarrow{d} Z_3 \quad (17)$$

where Z_3 is a d -dimensional normal random vector with zero mean and covariance matrix $\{ \nu_{0,2} - C(\delta)/4 - D(\delta)/2 \} f_X(x)^{-1} I(\theta_1(x), \theta_2(x))^{-1}$ and

$$\begin{aligned} D(\delta) = & \nu_{0,2} - \frac{C(\delta)}{4} - \frac{1}{16} \left\{ 4(1 + \sqrt{2})C(\sqrt{2} - 1, \delta/2) + (3 + 2\sqrt{2})C(2 - \sqrt{2}, \delta/2) \right. \\ & \left. + 2C(\sqrt{2}, \delta/2) + 4(1 - \sqrt{2})C(\sqrt{2} + 1, \delta/2) + (3 - 2\sqrt{2})C(\sqrt{2} + 2, \delta/2) \right\}. \end{aligned}$$

It follows from (17) that the asymptotic mean squared error of $\tilde{\theta}^{(3)}(x)$ is

$$\begin{aligned} \text{AMSE}\{\tilde{\theta}^{(3)}(x)\} = & \{nhf_X(x)\}^{-1} \text{tr}\{I(\theta_1(x), \theta_2(x))^{-1}\} \left\{ \nu_{0,2} - \frac{C(\delta)}{4} - \frac{D(\delta)}{2} \right\} \\ & + \frac{1}{4} h^4 \nu_{2,1}^2 \{ (\theta_1''(x))^2 + (\theta_2''(x))^2 \}. \end{aligned} \quad (18)$$

Remark 5. The optimal bandwidth of $\tilde{\theta}^{(3)}(x)$ minimizing $\text{AMSE}\{\tilde{\theta}^{(3)}(x)\}$ in (18) is

$$h_3 = \{\nu_{0,2} - C(\delta)/4 - D(\delta)/2\}^{1/5} \nu_{0,2}^{-1/5} h_0 \quad (19)$$

giving the optimal AMSE of $\tilde{\theta}^{(3)}(x)$:

$$\text{AMSE}_3(x) = \{\nu_{0,2} - C(\delta)/4 - D(\delta)\}^{4/5} \nu_{0,2}^{-4/5} \text{AMSE}_0(x). \quad (20)$$

Hence the asymptotic relative efficiency of $\tilde{\theta}^{(3)}(x)$ compared to $\hat{\theta}(x)$ is

$$\text{eff}\{\tilde{\theta}^{(3)}(x), \hat{\theta}(x)\} = \{\nu_{0,2} - C(\delta)/4 - D(\delta)/2\}^{-4/5} \nu_{0,2}^{4/5} \geq 1. \quad (21)$$

For any kernel K , $0 \leq D(\delta) \leq \frac{5}{8}\nu_{0,2}$ for all $\delta > 0$. Comparing (15) and (21) we find that $\tilde{\theta}^{(3)}(x)$ has a better asymptotic efficiency compared to $\tilde{\theta}^{(j)}(x)$, $j = 1, 2$, and the improvement is significant.

Remark 6. The conclusions in Remarks 3, 4 and 5 are all based on the total mean squared error measure given in (6), (10) and (18). In some circumstances, the coordinatewise mean squared errors may be used to accommodate different accuracy requirements. Then the coordinatewise optimal bandwidths for the different estimators follow the same relations as in (12) and (19), and the componentwise relative efficiencies of our estimators compared to the local linear MLE remain as in the right-hand sides of (15) and (21).

Remark 7. Existing data-driven bandwidth selection rules for the local linear maximum likelihood estimator $\hat{\theta}(x)$ include the cross-validation method of Aerts and Claeskens (1997) and the grid search approach of Fan, et al. (1998). To implement our estimators, one can simply replace $\hat{\theta}(x)$ by our estimators in the above mentioned procedures.

Remark 8. For any $j = 1, 2, 3$, in our construction of our variance reduction estimator $\tilde{\theta}^{(j)}$, the same value of δ is applied to obtain all the d parameter estimators $\tilde{\theta}_1^{(j)}(x) \cdots, \tilde{\theta}_d^{(j)}(x)$. If the curvature in the parameter curves $\theta_1(x), \cdots, \theta_d(x)$ varies

largely from one to another, then it may be more preferable to implement different δ values for $\tilde{\theta}_1^{(j)}(x) \cdots, \tilde{\theta}_d^{(j)}(x)$ and that can be done coordinatewise.

3. Simulation study and real application

3.1 Simulation study. We consider the following two models in our simulation study.

Model A (extreme value distribution):

$$P(Y \leq y|X = x) = \exp \left[- \left\{ 1 + \gamma(x) \frac{y - \mu(x)}{\sigma(x)} \right\}_+^{-1/\gamma(x)} \right] \quad (22)$$

where $\theta(x) = (\gamma(x), \mu(x), \sigma(x))^T$, $\sigma(x) = 1 + x^2$, $\mu(x) = -1 + 2x$, $\gamma(x) = -0.2$ or 0 , $(u)_+ = u$ for positive u and $(u)_+ = 0$ otherwise and X is uniformly distributed on $[0, 1]$.

Model B (logistic regression):

$$P(Y = 1|X = x) = \frac{\exp\{\theta(x)\}}{1 + \exp\{\theta(x)\}} \quad \text{and} \quad P(Y = 0|X = x) = \frac{1}{1 + \exp\{\theta(x)\}},$$

where $\theta(x) = \theta_1(x) = 7\{\exp\{-(x+1)^2\} + \exp\{-(x-1)^2\}\} - 5.5$ or $\theta(x) = \theta_2(x) = 2 - x^2$ and X is uniformly distributed on $[0, 1]$.

Note that $\gamma(x)$, $\mu(x)$ and $\sigma(x)$ in Model A are called the shape, location and scale parameter curves, respectively. The reason why we consider $\gamma(x)$ being a constant is suggested by the real data application in the next section. Model B was considered by Fan, et al. (1998) as well.

We drew 400 random samples of size $n = 400$ and $n = 600$ from both Model A and Model B, and took

$$\delta = \delta(x) = \min \left\{ 1, \frac{x}{(1 + \sqrt{1/2})h}, \frac{1 - x}{(1 + \sqrt{1/2})h} \right\}$$

in computing the variance reduction estimates. The biweight kernel $K(u) = \frac{15}{16}(1 - u^2)^2 I(|u| \leq 1)$ was employed.

For Model A, we kept the bandwidth h at 0.15 for both $\hat{\theta}(x)$ and $\tilde{\theta}^{(3)}(x)$. For Model B, we employed the data-driven method in Fan, et al. (1998) to choose the optimal bandwidth. That is, we searched the optimal h from 0.1 to 0.4 by step 0.01 to minimize the median of the integrated squared errors of the local linear maximum likelihood estimate. Then, this optimal bandwidth was applied to both $\hat{\theta}(x)$ and $\tilde{\theta}^{(3)}(x)$. Under model B, we also experimented with $h = 0.15$ for both estimates.

One way to measure the relative performance of our variance reduction estimate $\tilde{\theta}^{(3)}(x)$ to the local linear maximum likelihood estimate $\hat{\theta}(x)$ on each sample is to compute the ratio of the integrated squared error (ISE) of the latter to that of the former. Table 1 reports the mean and standard deviation of the ISE ratios obtained from the 400 samples. From Table 1, we clearly observe the effectiveness of the variance reduction techniques, especially when a small bandwidth is employed. In addition, the fact that, in all cases considered, the relative performance stays roughly the same when n changed from 400 to 600 indicates that $\tilde{\theta}^{(3)}(x)$ already achieves the asymptotic relative efficiency at moderate sample sizes.

3.2 Real application. We analyze the annual maximum temperatures (degrees Celsius) measured at Station De Bilt, the Netherlands from January 1, 1901 to December 31, 2003, say $y_1, \dots, y_{103}$; see Figure 1. Here $x_i = 1900 + i$ was transformed to $x_i = i/104$ for $i = 1, \dots, 103$. This data set is constructed by taking the maximum of daily maximum temperatures available at http://www.knmi.nl/voorl/kd/li-jsten/daggem/etmgeg_downl.cgi?language=eng.

First we applied the extreme value distribution (22) to this data set with $\gamma(x)$ and $\sigma(x)$ being constants. The setup for estimation is the same as in the simulation study except the choice of bandwidth. Here the cross-validation bandwidth method

in Aerts and Claeskens (1997) was employed. More specifically, we first computed $CV(h) = \sum_{i=1}^n \log f(Y_i; \hat{\theta}^{[l]}(x_i))$ for $h = 0.1, 0.101, 0.102, \dots, 0.4$, where $\hat{\theta}^{[l]}(x)$ is the local linear maximum likelihood estimate without the observation (x_i, y_i) , and then chose h to minimize the quantity $CV(h)$. Figure 2 depicts $CV(h)$ against h . Based on Figure 2, we employed $h = 0.165$ and $h = 0.191$, which correspond to the largest two values of h where local minima of $CV(h)$ occur, to compute the local linear maximum likelihood estimates and the corresponding variance reduction estimates $\tilde{\theta}^{(3)}(x)$; see Figures 3 and 4. In any of the cases, our curve estimate has less fluctuations than the original local linear estimate while the two suggest similar patterns of the parameter curves. And the results from $h = 0.191$ are more useful since the curve estimates are smoother.

To estimate the variances of these estimates, a bootstrapping approach similar to Davison and Ramesh (2000) was employed. That is, we take with replacement 1000 bootstrapping samples $\{\epsilon_{1,j}^*, \dots, \epsilon_{103,j}^*\}_{j=1}^{1000}$ from $\{(1 + \hat{\gamma}(x_i)(y_i - \hat{\mu}(x_i))/\hat{\sigma}(x_i))^{-1/\hat{\gamma}(x_i)}, i = 1, \dots, 103\}$. For each of $j = 1, \dots, 1000$, form a bootstrap sample by setting $y_{i,j}^* = \hat{\mu}(x_i) + \hat{\sigma}(x_i)\{(\epsilon_{i,j}^*)^{-\hat{\gamma}(x_i)} - 1\}/\hat{\gamma}(x_i)$, $i = 1, \dots, 103$ and then recalculate the local linear maximum likelihood estimator $\hat{\theta}(x_i)$ and the variance reduction estimator $\tilde{\theta}^{(3)}(x_i)$ based on the bootstrap sample $(x_1, y_{1,j}^*), \dots, (x_{103}, y_{103,j}^*)$ to give bootstrap estimates. Finally the bootstrap estimates of variances of $(\hat{\gamma}(x_i), \hat{\mu}(x_i), \hat{\sigma}(x_i))^T$ and $(\tilde{\gamma}^{(3)}(x_i), \tilde{\mu}^{(3)}(x_i), \tilde{\sigma}^{(3)}(x_i))^T$ are obtained as the respective sample variances of the 1000 bootstrap estimates. They are depicted in Figures 5 and 6. We observe from Figures 5 and 6 that, for each of the parameter curves, the variance reduction estimator has a substantially smaller bootstrap variance estimate than the local linear maximum likelihood estimator for the interior points of x .

Second, we applied model (22) to this data set with all three parameters being functions of x . We took $h = 0.191$. As before, the curve estimates and the boot-

strapped variance estimates are plotted in Figures 7 and 8. Again, we observe that the variance reduction estimator has a substantially smaller bootstrap variance estimate than the local linear maximum likelihood estimator for the interior points of x . Moreover, both Figures 7 and 8 indicate that γ and σ may be modeled as constants especially for interior points of x . From Figure 6 and Figure 8, we see that, for each of $\hat{\theta}(x)$ and $\tilde{\theta}^{(3)}(x)$, the bootstrap variance estimate of $\mu(x)$ is much smaller when γ and σ are treated as constants compared to the case where they depend on x .

Appendix: Proofs

Proof of Theorem 1. Put

$$\begin{aligned}
Q_1 &= \begin{pmatrix} 1 & 0 \\ 0 & \nu_{2,1} \end{pmatrix}, \quad Q_2 = \begin{pmatrix} \nu_{0,2} & \nu_{1,2} \\ \nu_{1,2} & \nu_{2,2} \end{pmatrix}, \quad Q_3 = \begin{pmatrix} 0 & \nu_{2,1} \\ \nu_{2,1} & 0 \end{pmatrix}, \\
\Sigma_x &= f_X(x) \begin{pmatrix} I_{11}(\theta_1(x), \theta_2(x))Q_1 & I_{12}(\theta_1(x), \theta_2(x))Q_1 \\ I_{21}(\theta_1(x), \theta_2(x))Q_1 & I_{22}(\theta_1(x), \theta_2(x))Q_1 \end{pmatrix}, \\
\Gamma_x &= f_X(x) \begin{pmatrix} I_{11}(\theta_1(x), \theta_2(x))Q_2 & I_{12}(\theta_1(x), \theta_2(x))Q_2 \\ I_{21}(\theta_1(x), \theta_2(x))Q_2 & I_{22}(\theta_1(x), \theta_2(x))Q_2 \end{pmatrix}, \\
\Lambda_x &= \begin{pmatrix} \frac{d}{dx}\{f_X(x)I_{11}(\theta_1(x), \theta_2(x))\}Q_3 & \frac{d}{dx}\{f_X(x)I_{12}(\theta_1(x), \theta_2(x))\}Q_3 \\ \frac{d}{dx}\{f_X(x)I_{21}(\theta_1(x), \theta_2(x))\}Q_3 & \frac{d}{dx}\{f_X(x)I_{22}(\theta_1(x), \theta_2(x))\}Q_3 \end{pmatrix}, \\
V_n(x) &= \sqrt{nh} \left(\hat{\theta}_1(x) - \theta_1(x), h\{\hat{\theta}'_1(x) - \theta'_1(x)\}, \hat{\theta}_2(x) - \theta_2(x), h\{\hat{\theta}'_2(x) - \theta'_2(x)\} \right)^T, \\
q_1(y; u_1, u_2) &= \frac{\partial}{\partial s} \log f(y; s, t)|_{(s,t)=(u_1, u_2)}, \\
q_2(y; u_1, u_2) &= \frac{\partial}{\partial t} \log f(y; s, t)|_{(s,t)=(u_1, u_2)}, \\
W_n(x) &= (W_{n1}(x), \dots, W_{n4}(x))^T, \\
W_{n(2k+l-1)}(x) &= \frac{\sqrt{h}}{\sqrt{n}} \sum_{i=1}^n (X - x_i)^l K_h(X_i - x) \\
&\quad q_k(Y_i; \theta_1(x) + \theta'_1(x)(X_i - x), \theta_2(x) + \theta'_2(x)(X_i - x)),
\end{aligned}$$

where $k = 1, 2$, $l = 0, 1$. Then it follows from Aerts and Claeskens (1997) that

$$(\Sigma_x + h\Lambda_x)V_n(x) - E\{W_n(x)\} = W_n(x) - E\{W_n(x)\} + o_p(1), \quad (23)$$

$$E\{W_n(x)\} = \sqrt{nh} \begin{pmatrix} \frac{1}{2}h^2 f(x)\nu_{2,1} \sum_{j=1,2} I_{1j}(\theta_1(x), \theta_2(x))\theta_j''(x) \{1 + o(1)\} \\ o(h^2) \\ \frac{1}{2}h^2 f(x)\nu_{2,1} \sum_{j=1,2} I_{2j}(\theta_1(x), \theta_2(x))\theta_j''(x) \{1 + o(1)\} \\ o(h^2) \end{pmatrix}. \quad (24)$$

Define

$$V_n^*(x) = \text{diag}(1, h, 1, h)\sqrt{nh}\{\tilde{\theta}^*(x) - \theta^*(x)\},$$

$$\tilde{\theta}^*(x) = \sum_{j=0,1,2} A_j(r)\hat{\theta}^*(\beta_{x,j}),$$

where $A_0(r) = 2^{-1}r(r-1)$, $A_1(r) = (1-r^2)$ and $A_2(r) = 2^{-1}r(1+r)$. Note that

$$\tilde{\theta}^*(x) - \theta^*(x) = \sum_{j=0,1,2} A_j(r)\{\hat{\theta}^*(\beta_{x,j}) - \theta^*(\beta_{x,j})\} + A_j(r)\{\theta^*(\beta_{x,j}) - \theta^*(x)\}, \quad (25)$$

$$1 = \sum_{j=0,1,2} A_j(r), \quad 0 = \sum_{j=0,1,2} (-1+j-r)A_j(r), \quad 0 = \sum_{j=0,1,2} (-1+j-r)^2 A_j(r). \quad (26)$$

Define

$$\begin{aligned} C_1^*(a, b) &= \int K(s + a\delta h)K(s + b\delta h) ds, \\ C_2^*(a, b) &= \int K(s + a\delta h)K(s + b\delta h)(s + a\delta h) ds, \\ C_3^*(a, b) &= \int K(s + a\delta h)K(s + b\delta h)(s + a\delta h)(s + b\delta h) ds, \\ \gamma_{ij} &= \text{Cov}\left\{ \sum_{l=0,1,2} A_l(r)W_{ni}(\beta_{x,l}) + \sum_{l=0,1,2} A_l(r)W_{nj}(\beta_{x,l}) \right\}. \end{aligned}$$

It is easy to check that $\gamma_{ij} = \gamma_{ji}$,

$$\gamma_{ll} = f_X(x)I_{11}(\theta_1(x), \theta_2(x)) \sum_{i=0,1,2} \sum_{j=0,1,2} A_i(r)A_j(r)C_{2l-1}^*(1-i+r, 1-j+r), \quad l = 1, 2,$$

$$\gamma_{12} = f_X(x)I_{11}(\theta_1(x), \theta_2(x)) \sum_{i=0,1,2} \sum_{j=0,1,2} A_i(r)A_j(r)C_2^*(1-i+r, 1-j+r),$$

$$(\gamma_{13}, \gamma_{14}, \gamma_{23}, \gamma_{24})^T = (\gamma_{11}, \gamma_{12}, \gamma_{12}, \gamma_{22})^T I_{12}(\theta_1(x), \theta_2(x)) / I_{11}(\theta_1(x), \theta_2(x)),$$

$$(\gamma_{33}, \gamma_{34}, \gamma_{44})^T = (\gamma_{11}, \gamma_{12}, \gamma_{22})^T I_{22}(\theta_1(x), \theta_2(x)) / I_{11}(\theta_1(x), \theta_2(x)).$$

From (24), (25) and (26), we have

$$\begin{aligned}
& \mathbb{E}\{(\Sigma_x + h\Lambda_x)V_n^*(x)\} \\
&= \sum_{j=0,1,2} A_j(r) \mathbb{E}\{W_n(\beta_{x,j})\} \{1 + o(1)\} \\
&\quad + \sum_{j=0,1,2} A_j(r) \text{diag}(1, h, 1, h) \sqrt{nh} \{\theta^*(\beta_{x,j}) - \theta^*(x)\} \\
&= \sum_{j=0,1,2} A_j(r) \mathbb{E}\{W_n(x)\} \{1 + o(1)\} \\
&\quad + \sum_{j=0,1,2} A_j(r) \text{diag}(1, h, 1, h) \sqrt{nh} \left(\theta'_1(x)(\beta_{x,j} - x) + \frac{1}{2} \theta''_1(x)(\beta_{x,j} - x)^2 + O(h^3), \right. \\
&\quad \left. \theta''_1(x)(\beta_{x,j} - x) + O(h^2), \theta'_2(x)(\beta_{x,j} - x) + \frac{1}{2} \theta''_2(x)(\beta_{x,j} - x)^2 + O(h^3), \right. \\
&\quad \left. \theta''_2(x)(\beta_{x,j} - x) + O(h^2) \right)^T \\
&= \mathbb{E}\{W_n(x)\} \{1 + o(1)\} + O(\sqrt{nh}h^3). \tag{28}
\end{aligned}$$

Similar to the proof of (23), we have (24) and (28) implying that

$$\begin{aligned}
& (\Sigma_x + h\Lambda_x) \tilde{V}_n^*(x) - \mathbb{E}\{W_n(x)\} \\
&= \sum_{j=0,1,2} A_j(r) [W_n(\beta_{x,j}) - \mathbb{E}\{W_n(\beta_{x,j})\}] \{1 + o_p(1)\} \\
&\xrightarrow{d} N(0, (\gamma_{ij})).
\end{aligned}$$

Hence, (4) can be shown by noting that

$$f_X(x) | I(\theta_1(x), \theta_2(x)) | \Sigma_x^{-1} = \begin{pmatrix} J_{22} & -J_{12} \\ -J_{21} & J_{11} \end{pmatrix}$$

where $J_{ij} = \text{diag}(I_{ij}(\theta_1(x), \theta_2(x)), \nu_{2,1}^{-1} I_{ij}(\theta_1(x), \theta_2(x)))$.

Proof of Theorem 2. It follows directly from Theorem 1.

Proof of Theorem 3. Similar to proof of Theorem 1.

References

- [1] Aerts, M., and Claeskens, G. (1997), “Local polynomial estimators in multiparameter likelihood models,” *Journal of the American Statistical Association*, 92, 1536-1545.
- [2] Beirlant, J., and Goegebeur, Y. (2004), “Local polynomial maximum likelihood estimation for Pareto-type distributions,” *Journal of Multivariate Analysis*, 89, 97-118.
- [3] Cheng, M.-Y., Peng, L., and Wu, J.-S. (2005), “Reducing variance in univariate smoothing,” *Technical report*.
- [4] Chavez-Demoulin, V., and Davison, A.C. (2005), “Generalized additive modelling of sample extremes,” *Journal of the Royal Statistical Society Series C*, 54, 207-222.
- [5] Claeskens, G., and Van Keilegom, I. (2003), “Bootstrap confidence bands for regression curves and their derivatives,” *The Annals of Statistics*, 31, 1852-1884.
- [6] Davison, A.C., and Ramesh, N.I. (2000), “Local likelihood smoothing of sample extremes,” *Journal of the Royal Statistical Society Series B*, 62, 191-208.
- [7] Eguchi, S., Kim, T.Y., and Park, B.U. (2003), “Local likelihood method: a bridge over parametric and nonparametric regression,” *Journal of Nonparametric Statistics*, 15, 665-683.
- [8] Fan, J., Farmen, M., and Gijbels, I. (1998), “Local maximum likelihood estimation and inference,” *Journal of the Royal Statistical Society Series B*, 60, 591-608.

- [9] Fan, J., Heckman, N.E., and 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.
- [10] Hall, P., and Tajvidi, N. (2000), “Nonparametric analysis of temporal trend when fitting parametric models to extreme-value data,” *Statistical Science*, 15, 153-167.
- [11] Irizarry, R.A. (2001), “Information and posterior probability criteria for model selection in local likelihood estimation,” *Journal of the American Statistical Association*, 96, 303-315.
- [12] Kogure, A. (1998), “Effective interpolations for kernel density estimators,” *Journal of Nonparametric Statistics*, 9, 165-195.
- [13] Loader, C. (1999), *Local regression and likelihood*, Springer.
- [14] Ramesh, N.I., and Davison, A.C. (2002), “Local models for exploratory analysis of hydrological extremes,” *Journal of Hydrology*, 256, 106-119.
- [15] Signorini, D.F., and Jones, M.C. (2004), “Kernel estimators for univariate binary regression,” *Journal of the American Association*, 99, 119-126.
- [16] Staniswalis, J.G. (1989), “The kernel estimate of a regression function in likelihood-based models,” *Journal of the American Statistical Association*, 84, 276-283.
- [17] Tibshirani, R., and Hastie, T. (1987), “Local likelihood estimation,” *Journal of the American Statistical Association*, 82, 559-567.
- [18] Yu, K., and Jones, M.C. (2004), “Likelihood-based local linear estimation of the conditional variance function,” *Journal of the American Statistical Association*, 99, 139-144.

Table 1. Relative performance. The numbers denote the mean of ratios of the integrated squared errors of the local linear maximum likelihood estimate $\hat{\theta}(x)$ to those of the variance reduction estimate $\tilde{\theta}^{(3)}(x)$ based on both $h = 0.15$ and the optimal h in the sense of minimizing the median of the integrated squared errors of the local linear maximum likelihood estimate. The corresponding standard deviations are given in brackets.

Model A				
h	$\gamma = -0.2, n = 400$	$\gamma = -0.2, n = 600$	$\gamma = 0.0, n = 400$	$\gamma = 0.0, n = 600$
0.15	γ : 1.394(0.504)	γ : 1.359(0.489)	γ : 1.476(0.488)	γ : 1.468(0.480)
	μ : 1.419(0.404)	μ : 1.411(0.425)	μ : 1.406(0.465)	μ : 1.406(0.457)
	σ : 1.384(0.455)	σ : 1.332(0.350)	σ : 1.395(0.405)	σ : 1.404(0.418)
Model B				
h	$\theta_1(x), n = 400$	$\theta_1(x), n = 600$	$\theta_2(x), n = 400$	$\theta_2(x), n = 600$
optimal	1.041(0.154)	1.027(0.152)	1.044(0.158)	1.052(0.157)
0.15	1.227(0.380)	1.227(0.372)	1.311(0.474)	1.281(0.416)

Captions for Figures

Caption for Figure 1. Annual maximum temperatures (degrees Celsius) measured at Station De Bilt, the Netherlands in the period of January 1, 1901 - December 31, 2003.

Caption for Figure 2. Cross-validation function for the annual maximum temperature data in the case of γ and σ being constants. The cross-validation function $CV(h)$ is plotted against h from 0.1 to 0.4 with step 0.001. The largest two values of h where local minima of $CV(h)$ occur are $h = 0.165$ and $h = 0.191$.

Caption for Figure 3. Estimators for the case of γ and σ being constants. The solid and dashed lines represent the local linear maximum likelihood estimator and the variance reduction estimator with bandwidth $h = 0.165$, respectively. Upper-left, upper-right and lower plots correspond to $\gamma(x)$, $\mu(x)$ and $\sigma(x)$.

Caption for Figure 4. Estimators for the case of γ and σ being constants. The solid and dashed lines represent the local linear maximum likelihood estimator and the variance reduction estimator with bandwidth $h = 0.191$, respectively. Upper-left, upper-right and lower plots correspond to $\gamma(x)$, $\mu(x)$ and $\sigma(x)$.

Caption for Figure 5. Bootstrapped variance estimates for the case of γ and σ being constants. The solid and dashed lines represent the bootstrapped variances of the local linear maximum likelihood estimator and the variance reduction estimator with bandwidth $h = 0.165$, respectively. Upper-left, upper-right and lower plots correspond to $\gamma(x)$, $\mu(x)$ and $\sigma(x)$.

Caption for Figure 6. Bootstrapped variance estimates for the case of γ and σ being constants. The solid and dashed lines represent the bootstrapped variances of the local linear maximum likelihood estimator and the variance reduction estimator with bandwidth $h = 0.191$, respectively. Upper-left, upper-right and lower plots

correspond to $\gamma(x)$, $\mu(x)$ and $\sigma(x)$.

Caption for Figure 7. Estimators for the case of all three parameters depending on x . The solid and dashed lines represent the local linear maximum likelihood estimator and the variance reduction estimator with bandwidth $h = 0.191$, respectively. Upper-left, upper-right and lower plots correspond to $\gamma(x)$, $\mu(x)$ and $\sigma(x)$.

Caption for Figure 8. Bootstrapped variance estimates for the case of all three parameters depending on x . The solid and dashed lines represent the bootstrapped variances of the local linear maximum likelihood estimator and the variance reduction estimator with bandwidth $h = 0.191$, respectively. Upper-left, upper-right and lower plots correspond to $\gamma(x)$, $\mu(x)$ and $\sigma(x)$.

Figure 1

Figure 2

Figure 3

Figure 4

Figure 5

Figure 6

Figure 7

Figure 8