Variance Reduction in Multiparameter Likelihood Models

Ming-Yen CHENG and Liang PENG

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, these models 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 article variance reduction techniques are employed to improve the 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; Variance reduction.

1. INTRODUCTION

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

$$f\big(y; \theta_1(x), \ldots, \theta_d(x)\big) \tag{1.1}$$

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 $\boldsymbol{\theta}(x) = (\theta_1(x), \ldots, \theta_d(x))^T$ depend on $X = x$. Under model (1.1), the dependence of Y on X is specified by a parametric law, with probability density function f, in which the parameter vector $\boldsymbol{\theta}$ is an unknown d-dimensional function of X.

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

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

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

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

Note that in model (1.1) what is essential is $\boldsymbol{\theta}(x) = (\theta_1(x), \ldots, \theta_d(x))^T$ and the derivatives $\boldsymbol{\alpha}(x) = (\theta_1'(x), \ldots, \theta_d'(x))^T$ are irrelevant. Hence, the local linear maximum likelihood estimator (MLE) for $\boldsymbol{\theta}(x)$ is defined as

$$\hat{\boldsymbol{\theta}}(x) = \big(\hat{\theta}_1(x), \hat{\theta}_2(x), \ldots, \hat{\theta}_d(x)\big)^T. \tag{1.2}$$

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

To provide a motivating example for studying variance reduction in the construction of multiparameter local likelihood models, we mention modeling of extremes and exceedances. Recently, there has been 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, Peng, and Wu (2005) introduced variance reduction techniques for nonparametric regression, which reduce the pointwise asymptotic variance uniformly. In this article we adopt the approach of Cheng et al. (2005) because 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): The multiparameter local likelihood model specifies the conditional distribution of the response variable given the covariates, whereas nonparametric regression considers the conditional mean, and asymptotic behaviors of nonparametric kernel regression are different from those of local likelihood estimation.

Ming-Yen Cheng is Professor, Department of Mathematics, National Taiwan University, Taipei 106, Taiwan (Email: *cheng@math.ntu.edu.tw*). Liang Peng is Associate Professor, School of Mathematics, Georgia Institute of Technology, Atlanta GA 30332 (Email: *peng@math.gatech.edu*). This article was supported in part by NSF grant DMS-04-03443 and a Humboldt research fellowship. The comments from a joint editor, an associate editor, and two reviewers are greatly appreciated.

© 2007 American Statistical Association
Journal of the American Statistical Association
March 2007, Vol. 102, No. 477, Theory and Methods
DOI 10.1198/016214506000000807

Here we discuss briefly the scope of local likelihood models outside extremes and exceedances. Local likelihood methods effectively model the 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 maximum likelihood estimator (MLE) reduces to the local linear regression estimator; see, for example, 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 (1.2) when $d = 1$ and applied it to local logistic regression and local partial likelihood estimation of Cox's proportional hazard model. Staniswalis (1989) considered a 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 a local Gamma model for survival analysis. Fan, Heckman, and Wand (1995) extended the idea to local quasi-likelihood estimation. The literature further includes Irizarry (2001), Eguchi, Kim, and Park (2003), and Signorini and Jones (2004).

This article is organized as follows. 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 (1.2). In addition, the main theoretical results and discussions on additional 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 of estimation 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, as in Cheng et al. (2005), a variance reduction estimator for $\boldsymbol{\theta}(x)$ is defined as

$$\tilde{\boldsymbol{\theta}}(x) = \frac{r(r-1)}{2}\hat{\boldsymbol{\theta}}(\beta_{x,0}) + (1-r^2)\hat{\boldsymbol{\theta}}(\beta_{x,1}) + \frac{r(r+1)}{2}\hat{\boldsymbol{\theta}}(\beta_{x,2}), \quad (2.1)$$

where $\hat{\boldsymbol{\theta}}(x) = (\hat{\theta}_1(x), \ldots, \hat{\theta}_d(x))^T$ is the local linear maximum likelihood estimate given in (1.2). If supp(X) were bounded, supp$(X) = [0,1]$, say, because $x - (1-r)\delta h = \beta_{x,0} < x < \beta_{x,2} = x + (1+r)\delta h$, then 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

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

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

Next we compare the asymptotic distributions of our variance reduction estimator $\tilde{\boldsymbol{\theta}}(x)$ and the local linear maximum likelihood estimator $\hat{\boldsymbol{\theta}}(x)$. 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 $\boldsymbol{\theta}(x) = (\theta_1(x), \theta_2(x))^T$ as

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

where

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

and 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 $\tilde{\boldsymbol{\theta}}(x)$.

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

$$\sqrt{nh}\left\{\tilde{\boldsymbol{\theta}}(x) - \boldsymbol{\theta}(x) - \frac{1}{2}h^2\nu_{2,1}\boldsymbol{\theta}''(x)\right\} \xrightarrow{d} \mathbf{Z}_1, \quad (2.2)$$

where $\boldsymbol{\theta}''(x) = (\theta_1''(x), \theta_2''(x))^T$ and $\mathbf{Z}_1$ is a d-dimensional normal random vector with mean $\mathbf{0}$ and covariance matrix $\{\nu_{0,2} - r^2(1-r^2)C(\delta)\}f_X(x)^{-1}\mathbf{I}(\theta_1(x),\theta_2(x))^{-1}$.

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

$$\sqrt{nh}\left\{\hat{\boldsymbol{\theta}}(x) - \boldsymbol{\theta}(x) - \frac{1}{2}h^2\nu_{2,1}\boldsymbol{\theta}''(x)\right\} \xrightarrow{d} \mathbf{Z}_2, \quad (2.3)$$

where $\mathbf{Z}_2$ is a d-dimensional normal random vector with mean $\mathbf{0}$ and covariance matrix $\nu_{0,2}f_X(x)^{-1}\mathbf{I}(\theta_1(x),\theta_2(x))^{-1}$.

Remark 1. When x is a boundary point, that is, x is close to the endpoints of supp(X), $\hat{\boldsymbol{\theta}}(x)$ and $\tilde{\boldsymbol{\theta}}(x)$ 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 because of a reduced number of data points there.

Define the asymptotic mean squared error (AMSE) of an estimator of $\boldsymbol{\theta}(x)$ as the sum of the trace of the covariance matrix and the squared norm of the asymptotic bias vector in its asymptotic Normal distribution. Then, from (2.2) and (2.3), the asymptotic mean squared errors of $\tilde{\boldsymbol{\theta}}(x)$ and $\hat{\boldsymbol{\theta}}(x)$ are, respectively,

$$\begin{aligned} \mathrm{AMSE}\{\tilde{\boldsymbol{\theta}}(x)\} &= \{nhf_X(x)\}^{-1}\mathrm{tr}\{\mathbf{I}(\theta_1(x),\theta_2(x))^{-1}\} \\ &\quad \times \{\nu_{0,2} - r^2(1-r^2)C(\delta)\} \\ &\quad + \frac{1}{4}h^4\nu_{2,1}^2\{(\theta_1''(x))^2 + (\theta_2''(x))^2\}, \quad (2.4) \end{aligned}$$

$$\text{AMSE}\{\hat{\boldsymbol\theta}(x)\} = \{nhf_X(x)\}^{-1}\text{tr}\big\{\mathbf{I}(\theta_1(x),\theta_2(x))^{-1}\big\}\nu_{0,2} + \frac{1}{4}h^4\nu_{2,1}^2\big\{(\theta_1''(x))^2 + (\theta_2''(x))^2\big\}. \quad (2.5)$$

Comparing (2.4) and (2.5), note that the asymptotic mean squared error of $\tilde{\boldsymbol\theta}(x)$ differs from that of $\hat{\boldsymbol\theta}(x)$ by the term $-r^2(1-r^2)C(\delta)\{nhf_X(x)\}^{-1}\text{tr}\{\mathbf{I}(\theta_1(x),\theta_2(x))^{-1}\}$. Note that $0 < r^2(1-r^2) \le 1/4$ 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, $0 \le C(\delta) \le 3\nu_{0,2}/2$ 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 (2.2) and (2.3), note that our variance reduction method simultaneously reduces asymptotic variances in estimating all the parameters $\theta_1(x),\dots,\theta_d(x)$ no matter what d, the number of parameters in the local likelihood model, is. This property holds for all our estimators, $\tilde{\boldsymbol\theta}^{(j)}(x), j = 1,2,3$, which are introduced in (2.6), (2.7), and (2.14) and are constructed based on $\tilde{\boldsymbol\theta}(x)$.

Note that r in the definition of $\tilde{\boldsymbol\theta}(x)$ 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 of what h, δ, and K are, and the resultant estimators are

$$\tilde{\boldsymbol\theta}^{(1)}(x) = \frac{1-2^{1/2}}{4}\hat{\boldsymbol\theta}\big(x-\big(1+2^{-1/2}\big)\delta h\big) + \frac{1}{2}\hat{\boldsymbol\theta}\big(x - 2^{-1/2}\delta h\big) + \frac{1+2^{1/2}}{4}\hat{\boldsymbol\theta}\big(x-\big(2^{-1/2}-1\big)\delta h\big) \quad (2.6)$$

and

$$\tilde{\boldsymbol\theta}^{(2)}(x) = \frac{1+2^{1/2}}{4}\hat{\boldsymbol\theta}\big(x+\big(2^{-1/2}-1\big)\delta h\big) + \frac{1}{2}\hat{\boldsymbol\theta}\big(x + 2^{-1/2}\delta h\big) + \frac{1-2^{1/2}}{4}\hat{\boldsymbol\theta}\big(x+\big(2^{-1/2}+1\big)\delta h\big). \quad (2.7)$$

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

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

$$\text{AMSE}\big\{\tilde{\boldsymbol\theta}^{(j)}(x)\big\} = \{nhf_X(x)\}^{-1}\text{tr}\big\{\mathbf{I}(\theta_1(x),\theta_2(x))^{-1}\big\} \times \left\{\nu_{0,2} - \frac{C(\delta)}{4}\right\} + \frac{1}{4}h^4\nu_{2,1}^2\big\{(\theta_1''(x))^2 + (\theta_2''(x))^2\big\}, \quad (2.8)$$

$j = 1,2.$

Remark 3. It follows from (2.5) that the optimal bandwidth minimizing $\text{AMSE}\{\hat{\boldsymbol\theta}(x)\}$ is

$$h_0 = \left\{\frac{\nu_{0,2}}{f_X(x)\nu_{2,1}^2}\right\}^{1/5}\left\{\frac{\text{tr}(\mathbf{I}(\theta_1(x),\theta_2(x))^{-1})}{(\theta_1''(x))^2 + (\theta_2''(x))^2}\right\}^{1/5} n^{-1/5}. \quad (2.9)$$

Similarly, for $j = 1,2$, the optimal bandwidth minimizing $\text{AMSE}\{\tilde{\boldsymbol\theta}^{(j)}(x)\}$ given in (2.8) is

$$h_j = \left\{\nu_{0,2} - \frac{C(\delta)}{4}\right\}^{1/5}\nu_{0,2}^{-1/5}h_0. \quad (2.10)$$

Remark 4. The bandwidth h_0 given in (2.9) yields the optimal AMSE of $\hat{\boldsymbol\theta}(x)$:

$$\text{AMSE}_0(x) = \left\{\frac{\nu_{0,2}\text{tr}(\mathbf{I}(\theta_1(x),\theta_2(x))^{-1})}{f_X(x)}\right\}^{4/5} \times \big\{\nu_{2,1}^2\big[(\theta_1''(x))^2 + (\theta_2''(x))^2\big]\big\}^{1/5}n^{-4/5}. \quad (2.11)$$

For $j = 1,2$, the optimal bandwidth given in (2.10) yields the optimal AMSE of $\tilde{\boldsymbol\theta}^{(j)}(x)$:

$$\text{AMSE}_j(x) = \left\{\nu_{0,2} - \frac{C(\delta)}{4}\right\}^{4/5}\nu_{0,2}^{-4/5}\text{AMSE}_0(x), \quad (2.12)$$

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

$$\text{eff}\big\{\tilde{\boldsymbol\theta}^{(j)}(x), \hat{\boldsymbol\theta}(x)\big\} = \left\{\nu_{0,2} - \frac{C(\delta)}{4}\right\}^{-4/5}\nu_{0,2}^{4/5} \ge 1. \quad (2.13)$$

Although theoretical results imply that more variance reduction is achieved by implementing larger δ values, that may introduce large finite-sample bias effects. We suggest taking $\delta = 1$ for general purposes. Slightly larger values of δ, for example, $\delta = 1.2$, 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 $\tilde{\boldsymbol\theta}^{(1)}(\cdot)$ and $\tilde{\boldsymbol\theta}^{(2)}(\cdot)$ from $\hat{\boldsymbol\theta}(\cdot)$ when using different δ choices.

Either of the variance reduction estimator $\tilde{\boldsymbol\theta}^{(1)}(x)$ or $\tilde{\boldsymbol\theta}^{(2)}(x)$ uses more information from data points on one side of x than those on the other side; see (2.6) and (2.7). One way to cancel out these finite-sample biases is to take the average of the two estimators

$$\tilde{\boldsymbol\theta}^{(3)}(x) = \frac{1}{2}\big\{\tilde{\boldsymbol\theta}^{(1)}(x) + \tilde{\boldsymbol\theta}^{(2)}(x)\big\}. \quad (2.14)$$

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

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

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

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

$$\sqrt{nh}\left\{\tilde{\boldsymbol\theta}^{(3)}(x) - \boldsymbol\theta(x) - \frac{1}{2}h^2\nu_{2,1}\boldsymbol\theta''(x)\right\} \xrightarrow{d} \mathbf{Z}_3, \quad (2.15)$$

where $\mathbf{Z}_3$ is a d-dimensional normal random vector with mean $\mathbf{0}$ and covariance matrix $\{\nu_{0,2} - C(\delta)/4 - D(\delta)/2\} \times f_X(x)^{-1}\mathbf{I}(\theta_1(x),\theta_2(x))^{-1}$ and

$$D(\delta) = \nu_{0,2} - \frac{C(\delta)}{4} - \frac{1}{16}\big\{4(1+\sqrt{2})C(\sqrt{2}-1,\delta/2) + (3+2\sqrt{2})C(2-\sqrt{2},\delta/2)$$

$$+ 2C(\sqrt{2}, \delta/2) + 4(1-\sqrt{2})C(\sqrt{2}+1, \delta/2) + (3-2\sqrt{2})C(\sqrt{2}+2, \delta/2)\bigr\}.$$

It follows from (2.15) that the asymptotic mean squared error of $\tilde{\boldsymbol\theta}^{(3)}(x)$ is

$$\mathrm{AMSE}\bigl\{\tilde{\boldsymbol\theta}^{(3)}(x)\bigr\} = \{nhf_X(x)\}^{-1}\mathrm{tr}\bigl\{\mathbf{I}(\theta_1(x), \theta_2(x))^{-1}\bigr\} \times \left\{\nu_{0,2} - \frac{C(\delta)}{4} - \frac{D(\delta)}{2}\right\} + \frac{1}{4}h^4\nu_{2,1}^2\bigl\{(\theta_1''(x))^2 + (\theta_2''(x))^2\bigr\}. \quad (2.16)$$

Remark 5. The optimal bandwidth of $\tilde{\boldsymbol\theta}^{(3)}(x)$ minimizing AMSE$\{\tilde{\boldsymbol\theta}^{(3)}(x)\}$ in (2.16) is

$$h_3 = \left\{\nu_{0,2} - \frac{C(\delta)}{4} - \frac{D(\delta)}{2}\right\}^{1/5}\nu_{0,2}^{-1/5}h_0, \quad (2.17)$$

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

$$\mathrm{AMSE}_3(x) = \left\{\nu_{0,2} - \frac{C(\delta)}{4} - D(\delta)\right\}^{4/5}\nu_{0,2}^{-4/5}\mathrm{AMSE}_0(x). \quad (2.18)$$

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

$$\mathrm{eff}\bigl\{\tilde{\boldsymbol\theta}^{(3)}(x), \hat{\boldsymbol\theta}(x)\bigr\} = \left\{\nu_{0,2} - \frac{C(\delta)}{4} - \frac{D(\delta)}{2}\right\}^{-4/5}\nu_{0,2}^{4/5} \ge 1. \quad (2.19)$$

For any kernel K, $0 \le D(\delta) \le \frac{5}{8}\nu_{0,2}$ for all $\delta > 0$. Comparing (2.13) and (2.19), we find that $\tilde{\boldsymbol\theta}^{(3)}(x)$ has a better asymptotic efficiency compared to $\tilde{\boldsymbol\theta}^{(j)}(x), j = 1, 2$, and the improvement is significant.

Remark 6. The conclusions in Remarks 3–5 are all based on the total mean squared error measure given in (2.4), (2.8), and (2.16). 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 (2.10) and (2.17), and the componentwise relative efficiencies of our estimators compared to the local linear MLE remain as on the right sides of (2.13) and (2.19).

Remark 7. Existing data-driven bandwidth selection rules for the local linear maximum likelihood estimator $\hat{\boldsymbol\theta}(x)$ 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 $\hat{\boldsymbol\theta}(x)$ by our estimators in the previously mentioned procedures.

Remark 8. For any $j = 1, 2, 3$, in our construction of our variance reduction estimator $\tilde{\boldsymbol\theta}^{(j)}$, the same value of δ is applied to obtain all the d-parameter estimators $\tilde\theta_1^{(j)}(x), \dots, \tilde\theta_d^{(j)}(x)$. If the curvature in the parameter curves $\theta_1(x), \dots, \theta_d(x)$ varies largely from one to another, then it may be more preferable to implement different δ values for $\tilde\theta_1^{(j)}(x), \dots, \tilde\theta_d^{(j)}(x)$ 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).

$$P(Y \le y|X = x) = \exp\left[-\left\{1 + \gamma(x)\frac{y - \mu(x)}{\sigma(x)}\right\}_+^{-1/\gamma(x)}\right], \quad (3.1)$$

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

Model B (Logistic regression).

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

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

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 that we consider $\gamma(x)$ as 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 Models A and 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 $K(u) = \frac{15}{16}(1 - u^2)^2 I(|u| \le 1)$ was employed.

For Model A, we kept the bandwidth h at .15 for both $\hat{\boldsymbol\theta}(x)$ and $\tilde{\boldsymbol\theta}^{(3)}(x)$. For Model B, we employed the data-driven method of Fan et al. (1998) to choose the optimal bandwidth. That is, we searched the optimal h from .1 to .4 in increments of .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 $\hat{\boldsymbol\theta}(x)$ and $\tilde{\boldsymbol\theta}^{(3)}(x)$. Under Model B, we also experimented with $h = .15$ for both estimates.

One way to measure the relative performance of our variance reduction estimate $\tilde{\boldsymbol\theta}^{(3)}(x)$ to the local linear maximum likelihood estimate $\hat{\boldsymbol\theta}(x)$ 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 $\tilde{\boldsymbol\theta}^{(3)}(x)$ already achieves the asymptotic relative efficiency at moderate sample sizes.

Table 1. Relative Performance

		Model A		
h	$\gamma = -.2$, *n* = 400	$\gamma = -.2$, *n* = 600	$\gamma = .0$, *n* = 400	$\gamma = .0$, *n* = 600
.15	γ: $1.394_{(.504)}$ μ: $1.419_{(.404)}$ σ: $1.384_{(.455)}$	γ: $1.359_{(.489)}$ μ: $1.411_{(.425)}$ σ: $1.332_{(.350)}$	γ: $1.476_{(.488)}$ μ: $1.406_{(.465)}$ σ: $1.395_{(.405)}$	γ: $1.468_{(.480)}$ μ: $1.406_{(.457)}$ σ: $1.404_{(.418)}$
		Model B		
h	$\theta_1(x)$, *n* = 400	$\theta_1(x)$, *n* = 600	$\theta_2(x)$, *n* = 400	$\theta_2(x)$, *n* = 600
Optimal	$1.041_{(.154)}$	$1.027_{(.152)}$	$1.044_{(.158)}$	$1.052_{(.157)}$
.15	$1.227_{(.380)}$	$1.227_{(.372)}$	$1.311_{(.474)}$	$1.281_{(.416)}$

NOTE: 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 $h = .15$ 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.

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 $y_1, \ldots, y_{103}$; see Figure 1. Here $x_i = 1{,}900 + i$ was transformed to $x_i = i/104$ for $i = 1, \ldots, 103$. This dataset is constructed by taking the maximum of daily maximum temperatures available at *http://www.knmi.nl/voorl/kd /lijsten/daggem/etmgeg_downl.cgi?language=eng*.

First, we applied the extreme value distribution (3.1) to this dataset 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 of Aerts and Claeskens (1997) was employed. More specifically, we first computed $\mathrm{CV}(h) = \sum_{i=1}^{n} \log f(Y_i; \hat{\boldsymbol{\theta}}^{[i]}(x_i))$ for $h = .1, .101, .102, \ldots, .4$, where $\hat{\boldsymbol{\theta}}^{[i]}(x)$ is the local linear maximum likelihood estimate without the observation (x_i, y_i), and then chose h to minimize the quantity $\mathrm{CV}(h)$. Figure 2 depicts $\mathrm{CV}(h)$ versus h. Based on Figure 2, we employed $h = .165$ and $h = .191$, which correspond to the largest two values of h where local minima of $\mathrm{CV}(h)$ occur, to compute the local linear maximum likelihood estimates and the corresponding variance reduction estimates $\tilde{\boldsymbol{\theta}}^{(3)}(x)$; 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. The results from $h = .191$ are more useful because the curve estimates are smoother.

To estimate the variances of these estimates, a bootstrapping approach similar to that of Davison and Ramesh (2000) was employed. That is, we take with replacement 1,000 bootstrapping samples $\{\epsilon_{1,j}^*, \ldots, \epsilon_{103,j}^*\}_{j=1}^{1{,}000}$ from $\{(1 + \hat{\gamma}(x_i)(y_i - \hat{\mu}(x_i))/\hat{\sigma}(x_i))^{-1/\hat{\gamma}(x_i)}, i = 1, \ldots, 103\}$. For each of $j = 1, \ldots, 1{,}000$, form a bootstrap sample by setting $y_{i,j}^* = \hat{\mu}(x_i) + \hat{\sigma}(x_i)\{(\epsilon_{i,j}^*)^{-\hat{\gamma}(x_i)} - 1\}/\hat{\gamma}(x_i)$, $i = 1, \ldots, 103$, and then recalculate the local linear maximum likelihood estimator $\hat{\boldsymbol{\theta}}(x_i)$ and the variance reduction estimator $\tilde{\boldsymbol{\theta}}^{(3)}(x_i)$ based on the bootstrap sample $(x_1, y_{1,j}^*), \ldots, (x_{103}, y_{103,j}^*)$ to give bootstrap estimates. Finally, the bootstrap estimates of variances of

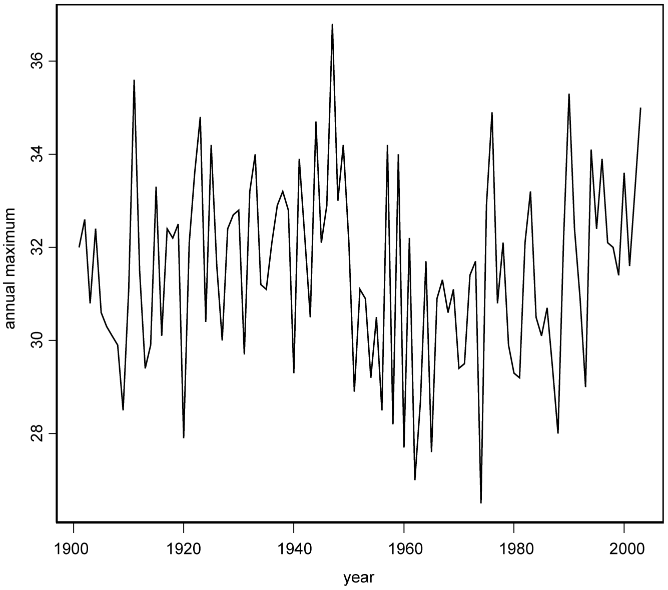

Figure 1. Annual Maximum Temperatures (degrees Celsius) Measured at Station De Bilt, the Netherlands, During the Period January 1, 1901–December 31, 2003.

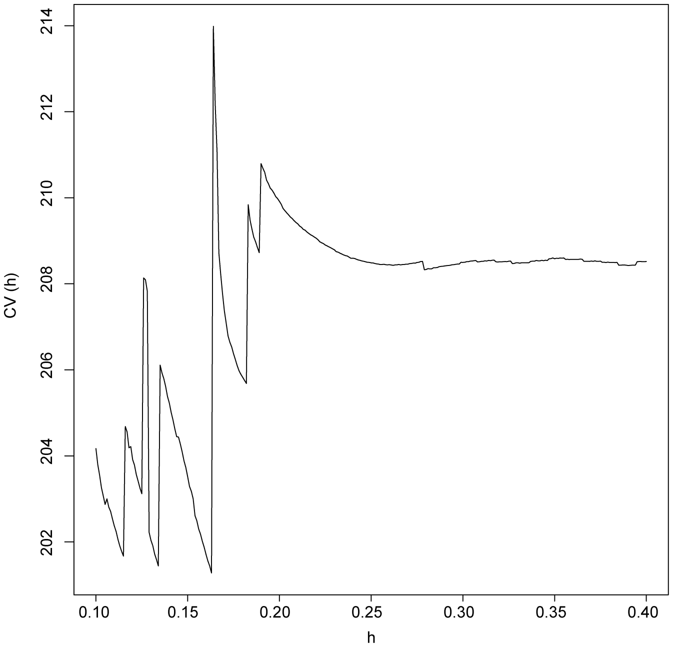

Figure 2. Cross-Validation Function for Annual Maximum Temperature Data When γ and σ Are Constants. The cross-validation function CV(h) is plotted against h from .1 to .4 with increments of .001. The largest two values of h where local minima of CV(h) occur are h = .165 and h = .191.

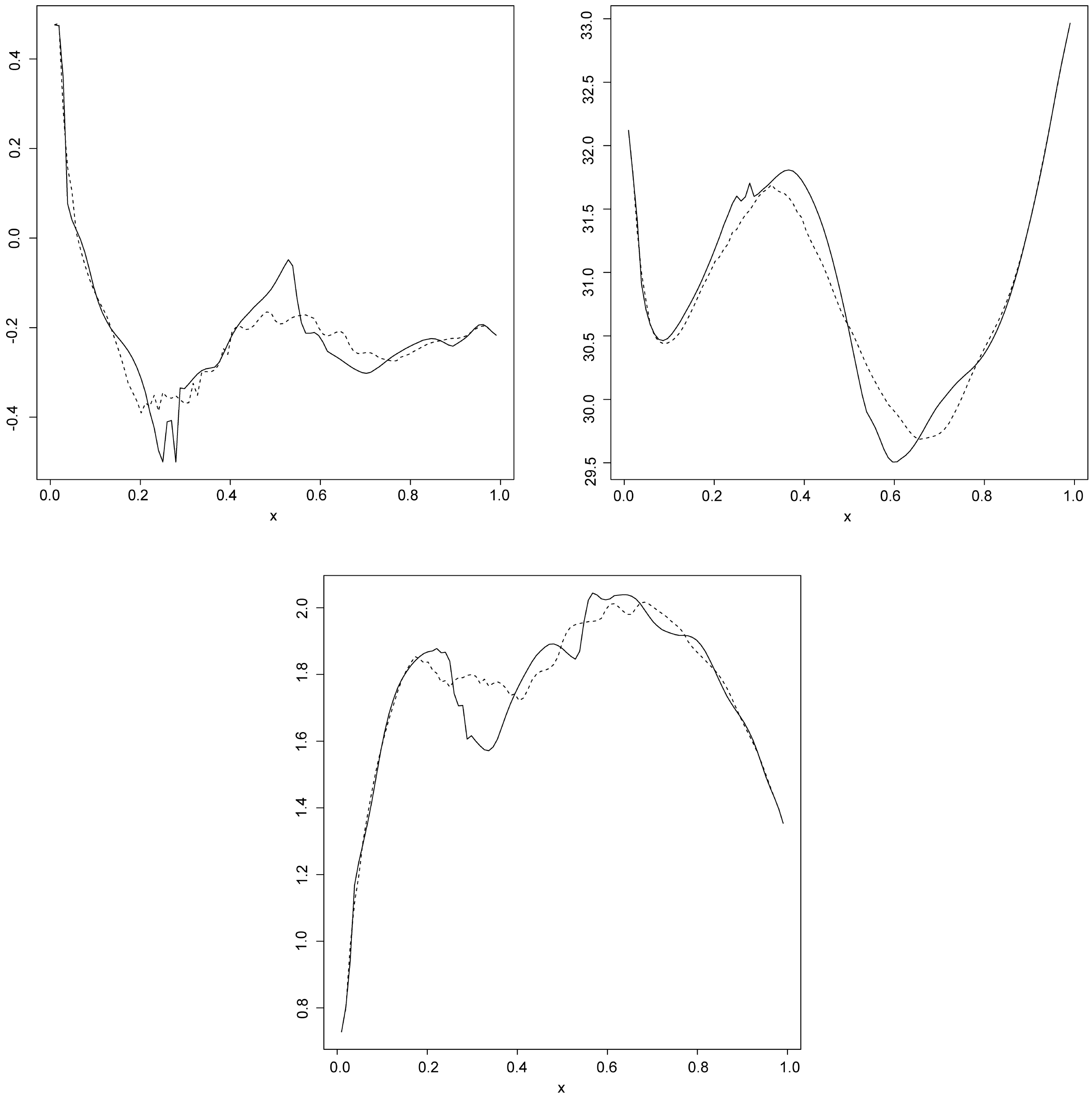

Figure 3. Estimators for the Case When γ and σ Are Constants. The solid and dashed lines represent the local linear maximum likelihood estimator and the variance reduction estimator with bandwidth $h = .165$, respectively. The upper left, upper right, and lower plots correspond to $\gamma(x)$, $\mu(x)$, and $\sigma(x)$, respectively.

$(\hat{\gamma}(x_i), \hat{\mu}(x_i), \hat{\sigma}(x_i))^T$ 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 1,000 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 (3.1) to this dataset with all three parameters being functions of x. We took $h = .191$. As before, the curve estimates and the bootstrapped 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 Figures 6 and 8, we see that, for each $\hat{\boldsymbol{\theta}}(x)$ and $\tilde{\boldsymbol{\theta}}^{(3)}(x)$, 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

$$\mathbf{Q}_1 = \begin{pmatrix} 1 & 0 \\ 0 & \nu_{2,1} \end{pmatrix}, \qquad \mathbf{Q}_2 = \begin{pmatrix} \nu_{0,2} & \nu_{1,2} \\ \nu_{1,2} & \nu_{2,2} \end{pmatrix},$$

$$\mathbf{Q}_3 = \begin{pmatrix} 0 & \nu_{2,1} \\ \nu_{2,1} & 0 \end{pmatrix},$$

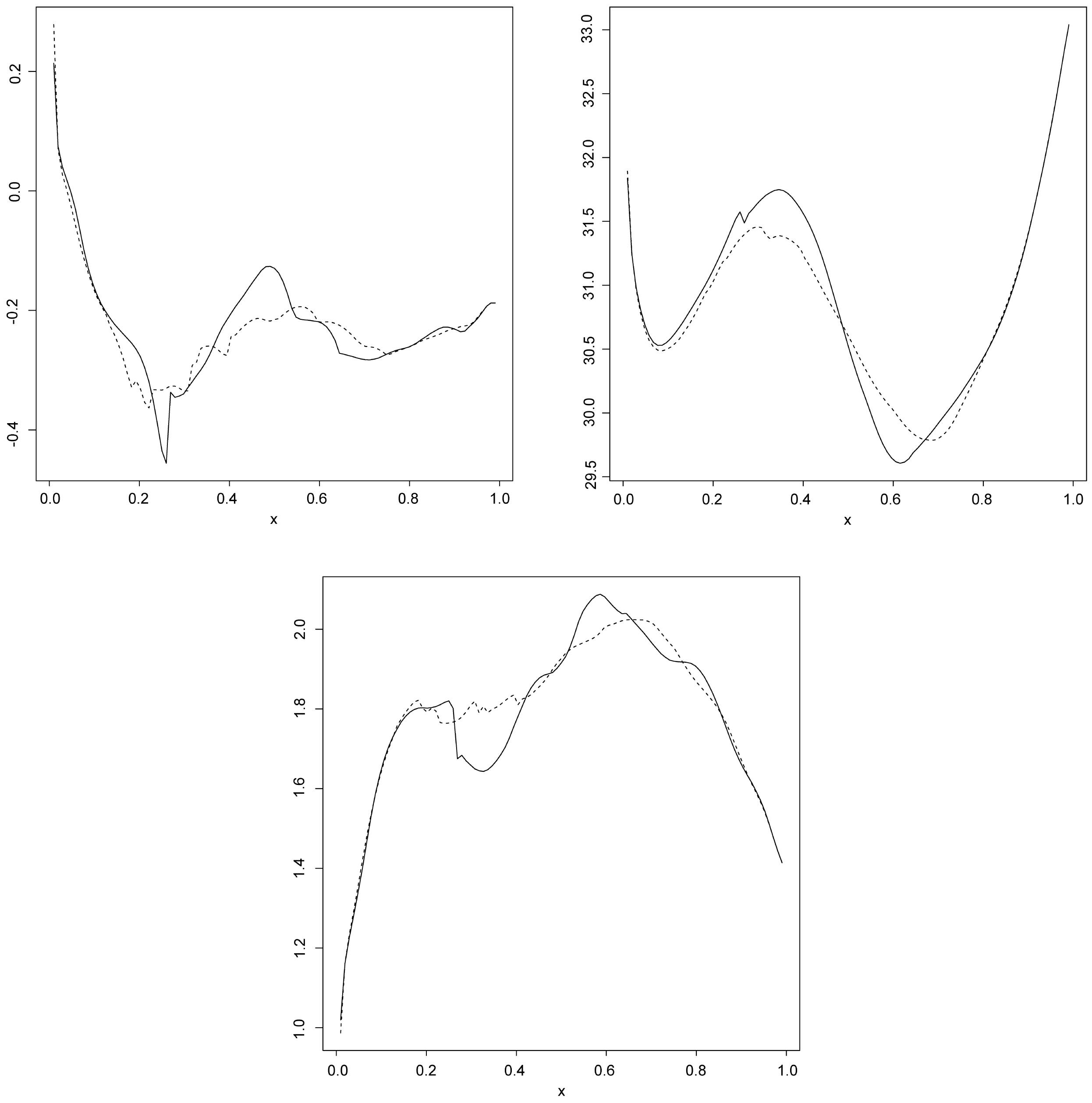

Figure 4. Estimators for the Case When γ and σ Are Constants. The solid and dashed lines represent the local linear maximum likelihood estimator and the variance reduction estimator with bandwidth $h = .191$, respectively. The upper left, upper right, and lower plots correspond to $\gamma(x)$, $\mu(x)$, and $\sigma(x)$, respectively.

$$\mathbf{\Sigma}_x = f_X(x)\begin{pmatrix} I_{11}(\theta_1(x),\theta_2(x))\mathbf{Q}_1 & I_{12}(\theta_1(x),\theta_2(x))\mathbf{Q}_1 \\ I_{21}(\theta_1(x),\theta_2(x))\mathbf{Q}_1 & I_{22}(\theta_1(x),\theta_2(x))\mathbf{Q}_1 \end{pmatrix},$$

$$\mathbf{\Gamma}_x = f_X(x)\begin{pmatrix} I_{11}(\theta_1(x),\theta_2(x))\mathbf{Q}_2 & I_{12}(\theta_1(x),\theta_2(x))\mathbf{Q}_2 \\ I_{21}(\theta_1(x),\theta_2(x))\mathbf{Q}_2 & I_{22}(\theta_1(x),\theta_2(x))\mathbf{Q}_2 \end{pmatrix},$$

$$\mathbf{\Lambda}_x = \begin{pmatrix} \frac{d}{dx}\{f_X(x)I_{11}(\theta_1(x),\theta_2(x))\}\mathbf{Q}_3 \\ \frac{d}{dx}\{f_X(x)I_{21}(\theta_1(x),\theta_2(x))\}\mathbf{Q}_3 \\ \frac{d}{dx}\{f_X(x)I_{12}(\theta_1(x),\theta_2(x))\}\mathbf{Q}_3 \\ \frac{d}{dx}\{f_X(x)I_{22}(\theta_1(x),\theta_2(x))\}\mathbf{Q}_3 \end{pmatrix},$$

$$\mathbf{V}_n(x) = \sqrt{nh}\big(\hat\theta_1(x)-\theta_1(x), h\{\hat\theta_1'(x)-\theta_1'(x)\}, \hat\theta_2(x)-\theta_2(x), h\{\hat\theta_2'(x)-\theta_2'(x)\}\big)^T,$$

$$q_1(y;u_1,u_2) = \frac{\partial}{\partial s}\log f(y;s,t)\Big|_{(s,t)=(u_1,u_2)},$$

$$q_2(y;u_1,u_2) = \frac{\partial}{\partial t}\log f(y;s,t)\Big|_{(s,t)=(u_1,u_2)},$$

$$\mathbf{W}_n(x) = \big(W_{n1}(x),\ldots,W_{n4}(x)\big)^T,$$

$$W_{n(2k+l-1)}(x) = \frac{\sqrt{h}}{\sqrt{n}}\sum_{i=1}^{n}(X-x_i)^l K_h(X_i-x) \times q_k\big(Y_i;\theta_1(x)+\theta_1'(x)(X_i-x), \theta_2(x)+\theta_2'(x)(X_i-x)\big),$$

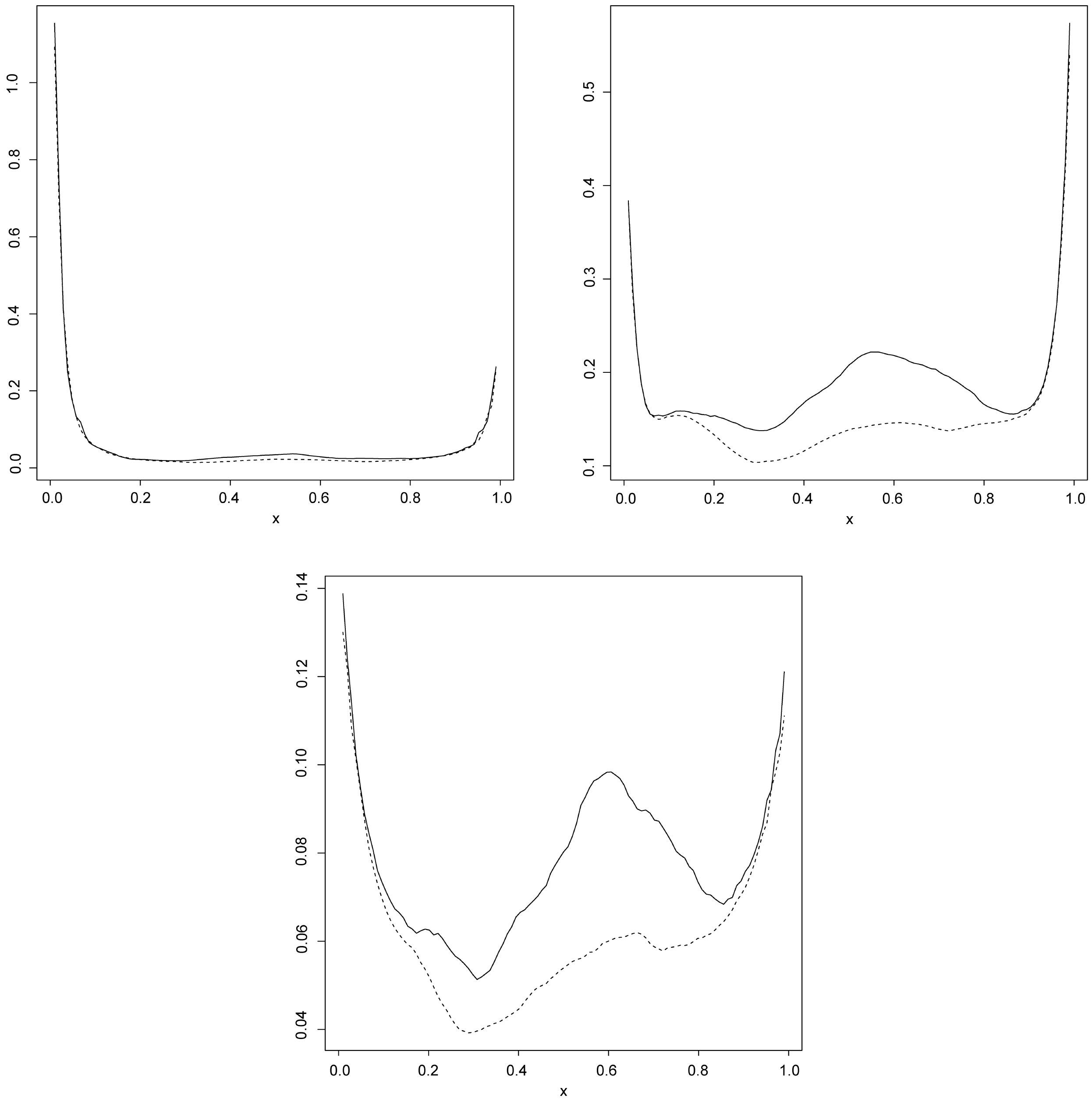

Figure 5. Bootstrapped Variance Estimates for the Case When γ and σ Are Constants. The solid and dashed lines represent the bootstrapped variances of the local linear maximum likelihood estimator and the variance reduction estimator with bandwidth $h = .165$, respectively. The upper left, upper right, and lower plots correspond to $\gamma(x)$, $\mu(x)$, and $\sigma(x)$, respectively.

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

$$(\boldsymbol{\Sigma}_x + h\boldsymbol{\Lambda}_x)\mathbf{V}_n(x) - \mathrm{E}\{\mathbf{W}_n(x)\} = \mathbf{W}_n(x) - \mathrm{E}\{\mathbf{W}_n(x)\} + o_p(1), \quad \text{(A.1)}$$

$$\mathrm{E}\{\mathbf{W}_n(x)\} = \sqrt{nh} \times \begin{pmatrix} \frac{1}{2}h^2 f(x)\nu_{2,1}\sum_{j=1,2} I_{1j}(\theta_1(x), \theta_2(x))\theta_j''(x)\{1 + o(1)\} \\ o(h^2) \\ \frac{1}{2}h^2 f(x)\nu_{2,1}\sum_{j=1,2} I_{2j}(\theta_1(x), \theta_2(x))\theta_j''(x)\{1 + o(1)\} \\ o(h^2) \end{pmatrix}. \quad \text{(A.2)}$$

Define

$$\mathbf{V}_n^*(x) = \mathrm{diag}(1, h, 1, h)\sqrt{nh}\{\tilde{\boldsymbol{\theta}}^*(x) - \boldsymbol{\theta}^*(x)\},$$
$$\tilde{\boldsymbol{\theta}}^*(x) = \sum_{j=0,1,2} A_j(r)\hat{\boldsymbol{\theta}}^*(\beta_{x,j}),$$

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

$$\tilde{\boldsymbol{\theta}}^*(x) - \boldsymbol{\theta}^*(x) = \sum_{j=0,1,2} A_j(r)\{\hat{\boldsymbol{\theta}}^*(\beta_{x,j}) - \boldsymbol{\theta}^*(\beta_{x,j})\} + A_j(r)\{\boldsymbol{\theta}^*(\beta_{x,j}) - \boldsymbol{\theta}^*(x)\}, \quad \text{(A.3)}$$

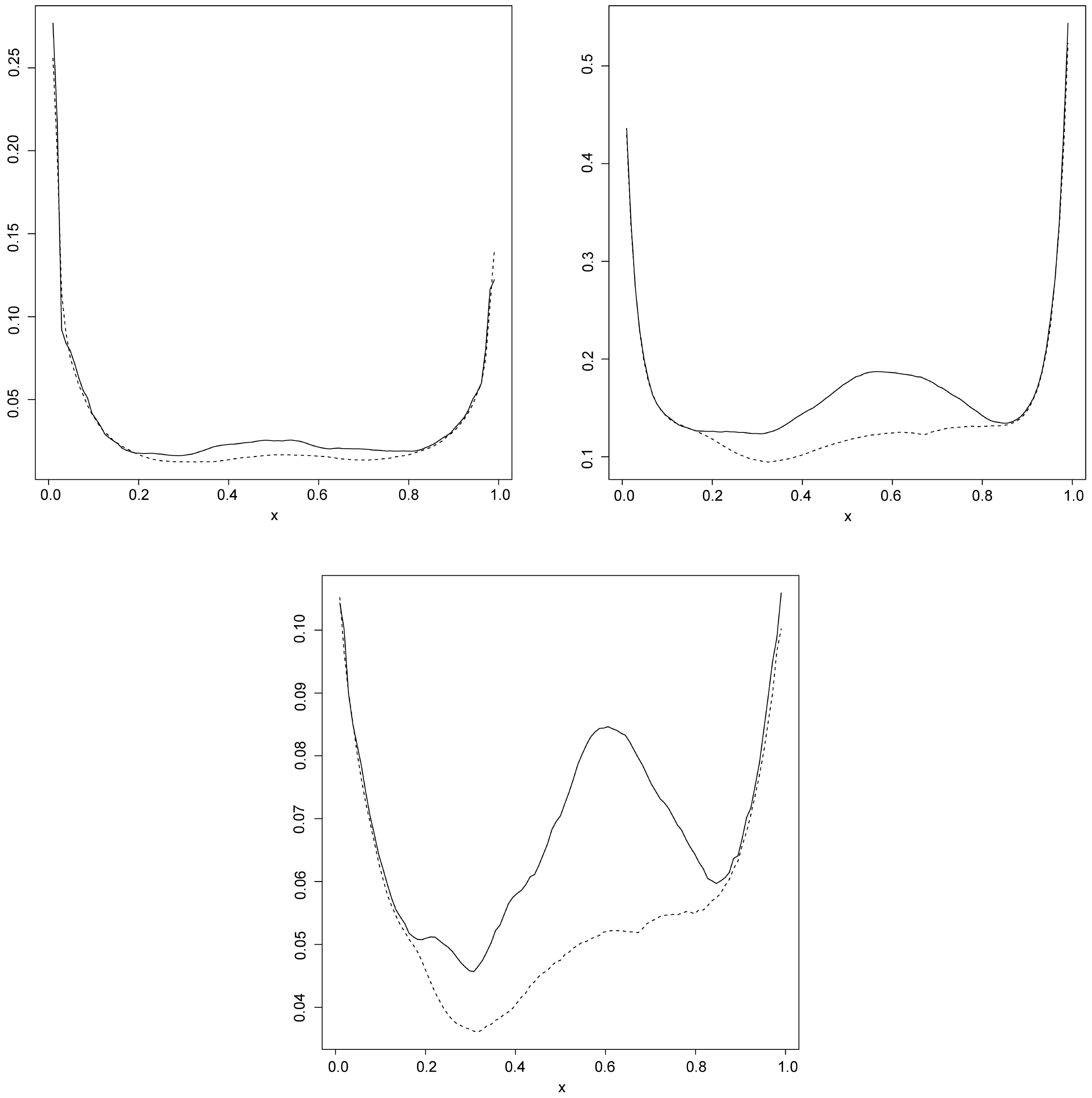

Figure 6. Bootstrapped Variance Estimates for the Case When γ and σ Are Constants. The solid and dashed lines represent the bootstrapped variances of the local linear maximum likelihood estimator and the variance reduction estimator with bandwidth $h = .191$, respectively. The upper left, upper right, and lower plots correspond to $\gamma(x)$, $\mu(x)$, and $\sigma(x)$, respectively.

$$
\begin{aligned}
1 &= \sum_{j=0,1,2} A_j(r), \\
0 &= \sum_{j=0,1,2} (-1+j-r)A_j(r), \\
0 &= \sum_{j=0,1,2} (-1+j-r)^2 A_j(r).
\end{aligned}
\tag{A.4}
$$

Define

$$C_1^*(a,b) = \int K(s+a\delta h)K(s+b\delta h)\,ds,$$

$$C_2^*(a,b) = \int K(s+a\delta h)K(s+b\delta h)(s+a\delta h)\,ds,$$

$$C_3^*(a,b) = \int K(s+a\delta h)K(s+b\delta h)(s+a\delta h)(s+b\delta h)\,ds,$$

$$\gamma_{ij} = \operatorname{cov}\left\{\sum_{l=0,1,2} A_l(r)W_{ni}(\beta_{x,l}),\ \sum_{l=0,1,2} A_l(r)W_{nj}(\beta_{x,l})\right\}.$$

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

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

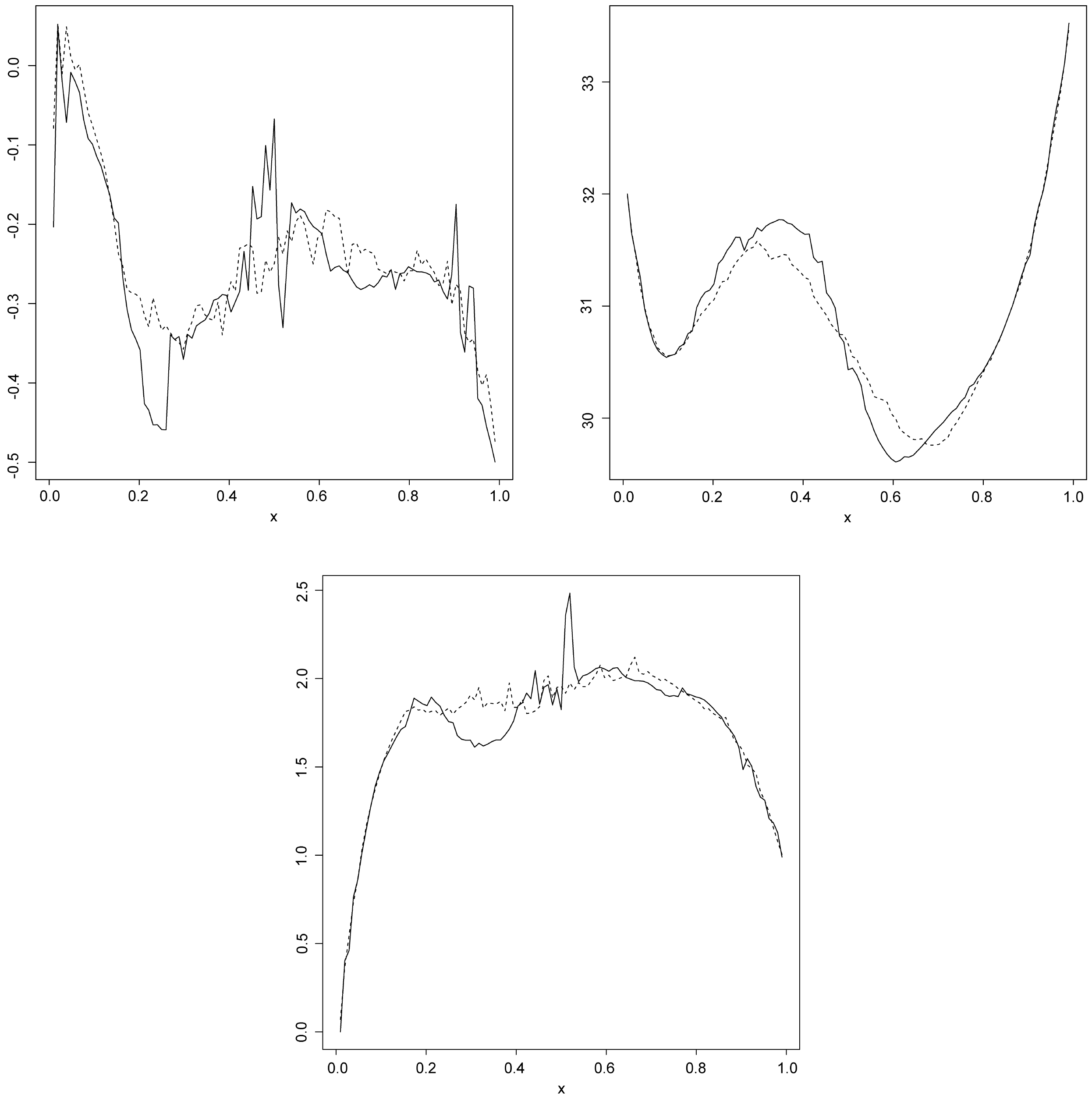

Figure 7. Estimators for the Case When All Three Parameters Depend on x. The solid and dashed lines represent the local linear maximum likelihood estimator and the variance reduction estimator with bandwidth h = .191, respectively. The upper left, upper right, and lower plots correspond to γ (x), μ (x), and σ (x), respectively.

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

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

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

From (A.2)–(A.4), we have

$$\begin{aligned}
&\mathrm{E}\{(\mathbf{\Sigma}_x + h\mathbf{\Lambda}_x)\mathbf{V}_n^*(x)\} \\
&\quad = \sum_{j=0,1,2} A_j(r)\mathrm{E}\{\mathbf{W}_n(\beta_{x,j})\}\{1 + o(1)\} \\
&\qquad + \sum_{j=0,1,2} A_j(r)\mathrm{diag}(1, h, 1, h)\sqrt{nh}\{\boldsymbol{\theta}^*(\beta_{x,j}) - \boldsymbol{\theta}^*(x)\} \\
&\quad = \sum_{j=0,1,2} A_j(r)\mathrm{E}\{\mathbf{W}_n(x)\}\{1 + o(1)\} \\
&\qquad + \sum_{j=0,1,2} A_j(r)\mathrm{diag}(1, h, 1, h)\sqrt{nh} \\
&\qquad \times \Bigg(\theta_1'(x)(\beta_{x,j} - x) + \frac{1}{2}\theta_1''(x)(\beta_{x,j} - x)^2 + O(h^3), \\
&\qquad\quad \theta_1''(x)(\beta_{x,j} - x) + O(h^2), \\
&\qquad\quad \theta_2'(x)(\beta_{x,j} - x) + \frac{1}{2}\theta_2''(x)(\beta_{x,j} - x)^2 + O(h^3),
\end{aligned}$$

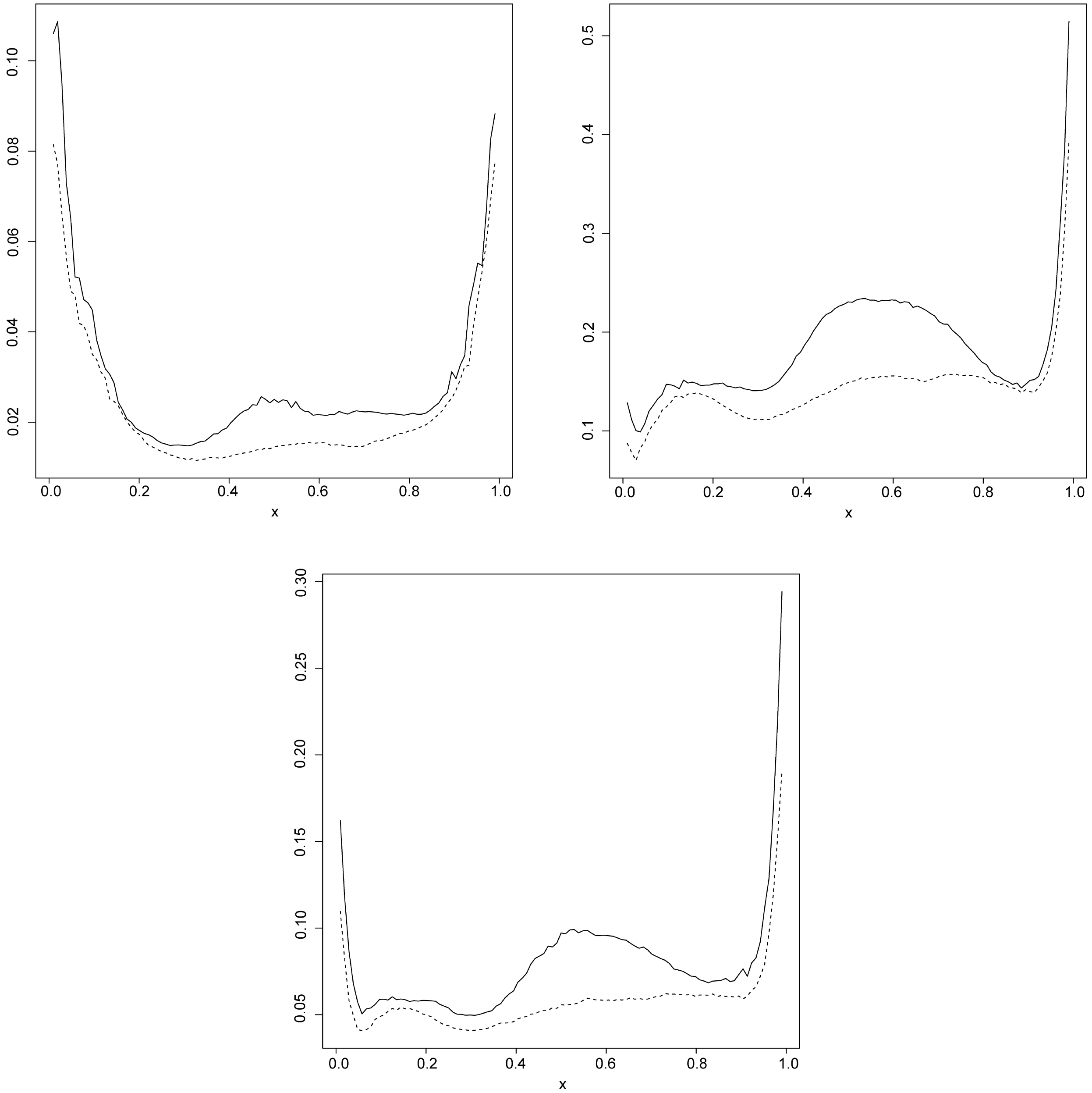

Figure 8. Bootstrapped Variance Estimates for the Case When All Three Parameters Depend on x. The solid and dashed lines represent the bootstrapped variances of the local linear maximum likelihood estimator and the variance reduction estimator with bandwidth $h = .191$, respectively. The upper left, upper right, and lower plots correspond to $\gamma(x)$, $\mu(x)$, and $\sigma(x)$, respectively.

$$\left.\theta_2''(x)(\beta_{x,j} - x) + O(h^2)\right)^T$$

$$= \mathrm{E}\{\mathbf{W}_n(x)\}\{1 + o(1)\} + O(\sqrt{nh}h^3). \qquad \text{(A.5)}$$

Similar to the proof of (A.1), we have (A.2), and (A.5) implying that

$$(\boldsymbol{\Sigma}_x + h\boldsymbol{\Lambda}_x)\tilde{\mathbf{V}}_n^*(x) - \mathrm{E}\{\mathbf{W}_n(x)\}$$
$$= \sum_{j=0,1,2} A_j(r)\big[\mathbf{W}_n(\beta_{x,j}) - \mathrm{E}\{\mathbf{W}_n(\beta_{x,j})\}\big]\{1 + o_p(1)\}$$
$$\xrightarrow{d} \mathrm{N}(0, (\gamma_{ij})).$$

Hence, (2.2) can be shown by noting that

$$f_X(x)\big|\mathbf{I}(\theta_1(x), \theta_2(x))\big|\boldsymbol{\Sigma}_x^{-1} = \begin{pmatrix} \mathbf{J}_{22} & -\mathbf{J}_{12} \\ -\mathbf{J}_{21} & \mathbf{J}_{11} \end{pmatrix},$$

where $\mathbf{J}_{ij} = \mathrm{diag}(I_{ij}(\theta_1(x), \theta_2(x)), \nu_{2,1}^{-1} I_{ij}(\theta_1(x), \theta_2(x)))$.

Proof of Theorem 2

Follows directly from Theorem 1.

Proof of Theorem 3

Similar to the proof of Theorem 1.

[Received May 2005. Revised May 2006.]

REFERENCES

Aerts, M., and Claeskens, G. (1997), "Local Polynomial Estimators in Multiparameter Likelihood Models," *Journal of the American Statistical Association*, 92, 1536–1545.

Beirlant, J., and Goegebeur, Y. (2004), "Local Polynomial Maximum Likelihood Estimation for Pareto-Type Distributions," *Journal of Multivariate Analysis*, 89, 97–118.

Chavez-Demoulin, V., and Davison, A. C. (2005), "Generalized Additive Modelling of Sample Extremes," *Journal of the Royal Statistical Society*, Ser. C, 54, 207–222.

Cheng, M.-Y., Peng, L., and Wu, J.-S. (2005), "Reducing Variance in Univariate Smoothing," technical report.

Claeskens, G., and Van Keilegom, I. (2003), "Bootstrap Confidence Bands for Regression Curves and Their Derivatives," *The Annals of Statistics*, 31, 1852–1884.

Davison, A. C., and Ramesh, N. I. (2000), "Local Likelihood Smoothing of Sample Extremes," *Journal of the Royal Statistical Society*, Ser. B, 62, 191–208.

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.

Fan, J., Farmen, M., and Gijbels, I. (1998), "Local Maximum Likelihood Estimation and Inference," *Journal of the Royal Statistical Society*, Ser. B, 60, 591–608.

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.

Hall, P., and Tajvidi, N. (2000), "Nonparametric Analysis of Temporal Trend When Fitting Parametric Models to Extreme-Value Data," *Statistical Science*, 15, 153–167.

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.

Kogure, A. (1998), "Effective Interpolations for Kernel Density Estimators," *Journal of Nonparametric Statistics*, 9, 165–195.

Loader, C. (1999), *Local Regression and Likelihood*, New York: Springer-Verlag.

Ramesh, N. I., and Davison, A. C. (2002), "Local Models for Exploratory Analysis of Hydrological Extremes," *Journal of Hydrology*, 256, 106–119.

Signorini, D. F., and Jones, M. C. (2004), "Kernel Estimators for Univariate Binary Regression," *Journal of the American Association*, 99, 119–126.

Staniswalis, J. G. (1989), "The Kernel Estimate of a Regression Function in Likelihood-Based Models," *Journal of the American Statistical Association*, 84, 276–283.

Tibshirani, R., and Hastie, T. (1987), "Local Likelihood Estimation," *Journal of the American Statistical Association*, 82, 559–567.

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.