ON LOCAL LIKELIHOOD DENSITY ESTIMATION
WHEN THE BANDWIDTH IS LARGE

Byeong U. Park\(^1\) and Young Kyung Lee\(^2\)

Department of Statistics, Seoul National University, Seoul, Korea

Tae Yoon Kim\(^3\) and Cheolyong Park\(^3\)

Department of Statistics, Keimyung University, Taegu, Korea

Shinto Eguchi\(^3\)

Institute of Statistical Mathematics, Tokyo, Japan

August 31, 2004: Final version for JSPI

Abstract

In this paper, we provide a large bandwidth analysis for a class of local likelihood methods. This work complements the small bandwidth analysis of Park, Kim and Jones (2002). Our treatment is more general than the large bandwidth analysis of Eguchi and Copas (1998). We provide a higher order asymptotic analysis for the risk of the local likelihood density estimator, from which a direct comparison between various versions of local likelihood can be made. The present work, being combined with the small bandwidth results of Park et al. (2002), gives an optimal size of the bandwidth which depends on the degree of departure of the underlying density from the proposed parametric model.


Key words and phrases. Density estimation, local likelihood, kernel function, bandwidth.

---

\(^1\)Research of Byeong U. Park was supported by KOSEF through Statistical Research Center for Complex Systems at Seoul National University.

\(^2\)Research of Young Kyung Lee was supported by the Brain Korea 21 project in 2004.

\(^3\)Research of Shinto Eguchi, Tae Yoon Kim and Cheolyong Park was supported by KOSEF(F01-2000-000-20002-0) and JSPS through Korea-Japan Joint Research Program.
1. Introduction

Local likelihood methods for density estimation are very promising, not least on account of their extraordinary flexibility and adaptivity. They afford efficient estimation if the proposed parametric family includes a good model for the data, as well as the usual good behaviour of nonparametric density estimation if it does not. Each formulation of local likelihood estimation is based on a locally weighted log-likelihood where the weights are determined by a kernel function $K$ and a bandwidth $h$. When $h$ is large, the resulting estimator is close to the fully parametric maximum likelihood estimator. On the other hand, when $h$ is small, performance of the resulting estimator would not depend much on the proposed parametric model. For consistent estimation, the methods require a correction term added to the naive locally weighted log-likelihood. Precise details of the correction term allow scope for variation. See Copas (1995), Loader (1996), Hjort and Jones (1996), and Kim, Park and Kim (2001) for statistical analysis on different formulations of local likelihood which employ a specific choice of the correction term.

Recently, Eguchi and Copas (1998) made an important contribution to local likelihood density estimation. They provided a unified formulation of various local likelihood approaches by introducing an arbitrary function, denoted by $\xi$, for the additional correction term. This unified formulation includes as special cases the C-version due to Copas (1995), the U-version due to Loader (1996) and Hjort and Jones (1996), and the T-version due to Eguchi and Copas (1998). The definition of the general local likelihood function and those of the U-, C-, and T-versions are given in the next section. Later, Park et al. (2002) presented a small $h$ asymptotics of the class of local likelihood estimators. It is based on a general condition on $\xi$ tailored for the small $h$ analysis. Eguchi and Copas (1998) found some interesting large $h$ properties of the class. However, their results are based on a rather stringent condition on $\xi$. The condition excludes some important local likelihood methods such as Copas (1995). See, for more details, the paragraph including (2.4) in the current paper, or the discussion in Eguchi and Copas (1998) immediately below the proof of their Theorem 1. Furthermore, their results are restricted to the case where the underlying density is ‘undetectably close’ to the parametric model.

In this paper, we give a more relevant large $h$ analysis with a quite general condition on the function $\xi$. The condition is different from the one used in Park et al. (2002) for a small $h$ analysis. The U-, C- and T-versions of local likelihood are seen to satisfy our new
condition. We provide a useful approximation of the risk when the bandwidth $h$ tends to infinity as sample size $n$ grows. We show that the mean integrated squared error takes the form
\[
C_{B,ML} + C_{V,ML} \frac{1}{n} - C_{B,SM} \frac{1}{h^2} - C_{V1,SM} \frac{1}{nh^2} + C_{V2,SM} \frac{1}{nh^4}.
\]
(1.1)
The first term $C_{B,ML}$ represents the model misspecification error of the proposed parametric family. This and $C_{V1,SM}$ are zero if the true density actually belongs to the parametric model. In fact, $C_{B,ML}$ and $n^{-1}C_{V,ML}$ are the integrated squared bias and the integrated variance, respectively, of the parametric maximum likelihood estimator. The next three terms are originated from local smoothing. The constants $C_{B,SM}$ and $C_{V1,SM}$ depend on the distance between the true density and the parametric model, while the last $C_{V2,SM}$ does not. Depending on how close the true density is to the parametric model, the last term $n^{-1}h^{-4}C_{V2,SM}$ may dominate $n^{-1}h^{-2}C_{V1,SM}$. We give explicit expressions for these constants. Our formula is more useful than the one given by Eguchi and Copas (1998) in the sense that its dependence on the choice of $\xi$ is more transparent, thus a direct comparison between various local likelihood methods can be made.

Let \( \{f(\cdot, \theta) : \theta \in \Theta\} \) be a parametric model proposed for the data where $\theta$ is a $p$-dimensional parameter vector. Let $\theta_\infty$ denote the solution of the population version of the parametric likelihood equation. In fact, $f(\cdot, \theta_\infty)$ minimizes the Kullback-Leibler divergence from the true density among all members in the parametric model. It is the best parametric approximant to the true density in that sense. A precise definition of $\theta_\infty$ is given at (2.8) in Section 2. We evaluate the risk in a shrinking neighborhood of the parametric model. In particular, we assume that the true density $g \equiv g_n$ satisfies
\[
\int \{g(x) - f(x, \theta_\infty)\}^2 dx \leq cn^{-(1+\alpha)},
\]
(1.2)
for some $\alpha > -1$ and $c > 0$. Note that $\alpha = \infty$ corresponds to the case where the true density belongs to the parametric model. We derive an approximation of the risk of the form given at (1.1) under this condition. It is shown that the constants $C_{B,ML}$ and $C_{B,SM}$ in (1.1) are both $O(n^{-1-\alpha})$. The constant $C_{V1,SM}$ is seen to be $O(n^{-(1+\alpha)/2})$.

Our large $h$ asymptotics is valid for all $\alpha > -1$. Note that, when $\alpha = -1$, a risk analysis in the usual nonparametric context of the bandwidth tending to zero as $n$ grows is more appropriate. Eguchi and Copas (1998) gave some interesting large $h$ properties of their class of local likelihood density estimators, but only in the case where $\alpha > 0$. 

3
Thus, the present work is also an extension of Eguchi and Copas (1998) in this regard. Since features of the underlying density (such as location and scale) may be estimated at best to an accuracy of $O(n^{-1/2})$, treatment of the case $-1 < \alpha \leq 0$, that is not dealt in Eguchi and Copas (1998), is useful to describe the large $h$ properties when the underlying density lies in a region where one can distinguish it from the parametric model. Our large $h$ asymptotics and the small $h$ results of Park et al. (2002), the latter being valid for all $\alpha \geq -1$, may be combined to give an optimal size of the bandwidth going to zero or infinity. It is argued that the value of $\alpha$ at which a transition from a large to a small bandwidth is desirable does exist in the range $-1 < \alpha \leq 0$. We observe in an example that the optimal bandwidth indeed ranges from a very small to a very large value depending on how close the true density is to the parametric model.

Our treatment of large bandwidth asymptotics is more rigorous than that of Eguchi and Copas (1998). We obtain higher order expansions for the bias and the variance of the local likelihood estimator. We show that by choosing the bandwidth in an optimal way the local likelihood estimator may have smaller risk than the parametric maximum likelihood estimator except the case where $\alpha = \infty$. The risk considered in this paper is the mean integrated weighted squared error. An expansion of the relative entropy risk (expected Kullback-Leibler divergence) considered in Eguchi and Copas (1998) is readily obtained from our results by choosing proper weight functions.

This paper is organized as follows. Section 2 introduces the unified formulation of local likelihood density estimation and provides the new condition on $\xi$ for large $h$ analysis. It also contains some preliminary results for the risk analysis. In Section 3, we give a detailed account of large $h$ asymptotics for all $\alpha > -1$. In Section 4, the risks of the U-, C- and T-versions of local likelihood estimation are compared with an example. Also discussed is the issue of optimal bandwidth size. Some technical proofs are given in the appendix.

2. Preliminaries

Let $X_1, \ldots, X_n$ be independent $d$-variate random vectors from a common density $g(\cdot)$ supported on $\mathcal{X}$. Let $f(\cdot, \theta)$ with $\theta$ being a $p$-dimensional parameter vector be a parametric model proposed for the data. Define $u(t, \theta) = (\partial/\partial \theta) \log \{ f(t, \theta) \}$. Let $K$ be a nonnegative symmetric kernel function on $\mathbb{R}^d$. For simplicity of presentation we take a
scalar bandwidth $h$. For an arbitrary function $\xi(\cdot, \cdot)$, the general form of local likelihood
estimating equation considered by Eguchi and Copas (1998) and Park et al. (2002) is given
by $\hat{\Psi}_h(x, \theta) = 0$ where

$$
\hat{\Psi}_h(x, \theta) = \frac{1}{n} \sum_{i=1}^{n} K\left(\frac{x - X_i}{h}\right) u(X_i, \theta) - \frac{1}{n} \sum_{i=1}^{n} \xi \left\{ K\left(\frac{x - X_i}{h}\right), E_{\theta} K\left(\frac{x - X_1}{h}\right) \right\} \\
x E_{\theta} \left\{ K\left(\frac{x - X_1}{h}\right) u(X_1, \theta) \right\}.
$$

Here and below $E_{\theta}$ means expectation with respect to the parametric density $f(\cdot, \theta)$
while $E$ with no subscript means expectation with respect to the true density $g$. If we
 take $h = \infty$, this reduces to the parametric maximum likelihood estimating equation:

$$
\sum_{i=1}^{n} u(X_i, \theta) = 0.
$$

Let $\hat{\theta}_h(x)$ denote the solution of the estimating equation $\hat{\Psi}_h(x, \theta) = 0$. A local
likelihood density estimator may be defined by

$$
\hat{f}_h(x) = f(x, \hat{\theta}_h(x)).
$$

(2.1)

Note that this estimator may not integrate to one since $\hat{\theta}_h(x)$ depends on $x$. A bona fide
density estimator is given by

$$
\hat{g}_h(x) = \hat{f}_h(x) / \int \hat{f}_h(x) \, dx.
$$

(2.2)

The function $\xi$ is assumed to satisfy

$$
E_{\theta} \xi \left\{ K\left(\frac{x - X_1}{h}\right), E_{\theta} K\left(\frac{x - X_1}{h}\right) \right\} = 1.
$$

(2.3)

This condition is required for the estimating equation to be unbiased when $g(\cdot) = f(\cdot, \theta)$
for some $\theta$. In fact, it guarantees the first-order Bartlett’s identity: $E_{\theta} \hat{\Psi}_h(x, \theta) = 0$ for
all $\theta$. Examples of $\xi$ that satisfy the condition (2.3) include the U-version $\xi(u, v) \equiv 1$ of
Hjort and Jones (1996), the C-version $\xi(u, v) = (1 - u)/(1 - v)$ of Copas (1995), and the
T-version $\xi(u, v) = u/v$ considered in Eguchi and Copas (1998).

Eguchi and Copas (1998) assumed $\alpha > 0$ and discussed the properties of $f(x, \hat{\theta}_h(x))$
when $h$ tends to infinity as $n$ grows. Their results are based on the following additional
condition on $\xi$:

$$
E \xi \left\{ K\left(\frac{x - X_1}{h}\right), E_{\theta_{\infty}} K\left(\frac{x - X_1}{h}\right) \right\} = 1 + O\left(\frac{1}{h^2}\right)
$$

(2.4)
as \( h \) tends to infinity. But, this condition is too restrictive. The C-version \( \xi(u, v) = (1 - u)/(1 - v) \), for example, does not satisfy this condition. To see this, assume as in Eguchi and Copas (1998) that \( K(t) = 1 - \kappa_2\|t\|^2 + O(\|t\|^4) \) as \( t \to 0 \). Then,

\[
E \xi \left\{ K \left( \frac{x - X_1}{h} \right), E_{\theta_{\infty}} K \left( \frac{x - X_1}{h} \right) \right\} \simeq 1 + \frac{1}{E_{\theta_{\infty}} \|x - X_1\|^2} \int \|x - y\|^2 \{ g(y) - f(y, \theta_{\infty}) \} \, dy.
\]

The second term in the above approximation is not \( O(h^{-2}) \), but equals \( O\{n^{-(1+\alpha)/2}\} \) when the true density \( g \equiv g_n \) satisfies the condition (1.2).

We consider here a more relevant condition on \( \xi \) which replaces (2.4). We note that, under the condition \( K(t) = 1 - \kappa_2\|t\|^2 + O(\|t\|^4) \) as \( t \to 0 \), the large \( h \) properties of the local likelihood estimator depend on \( \xi \) through the behaviour of \( \xi(1 - y, 1 - z) \) when both \( y \) and \( z \) approach to zero from above. This can be seen from the fact that both arguments of \( \xi \) in the definition of \( \hat{\Psi}_h \) converge to one as \( h \) tends to infinity. The condition on \( \xi \) should be different from the one in small \( h \) setting, where performance of the estimator relies on the properties of \( \xi(y, z) \) near \( y = z = 0 \) since both arguments of \( \xi \) tend to zero as \( h \) converges to zero. See Park et al. (2002) for a suitable condition on \( \xi \) in small \( h \) setting. The condition on \( \xi \) for our large \( h \) analysis is given in (A1) below. The condition on the kernel \( K \) is stated in (A2) where we specify the coefficient of \( O(\|t\|^4) \) term for more detailed analysis.

**Assumptions.**

(A1) In addition to the consistency condition (2.3), \( \xi \) satisfies

\[
\lim_{z \downarrow 0} \sup_{0 \leq y \leq c} \left| \frac{1}{z} \left\{ \xi(1 - yz, 1 - z) - \xi_0(y) - \xi_1(y)z \right\} \right| = 0 \tag{2.5}
\]

for some functions \( \xi_0, \xi_1 \) and a constant \( c > 0 \). The function \( \xi_0 \) is continuously differentiable and \( \xi_1 \) is continuous. Also, \( \xi(y, z) \) is twice continuously differentiable with respect to \( z \) on \((0, 1)\) for each \( y \in [0, 1] \).

(A2) The kernel function \( K(t) \) is continuous at \( t = 0 \) and satisfies \( K(t) = 1 - \kappa_2\|t\|^2 + \kappa_4\|t\|^4 + o(\|t\|^4) \) as \( \|t\| \to 0 \).

The three versions of local likelihood mentioned above satisfy the condition (A1): for the U-version, \( \xi_0(y) \equiv 1 \) and \( \xi_1(y) \equiv 0 \); for the C-version, \( \xi_0(y) = y \) and \( \xi_1(y) \equiv 0 \);
and for the T-version, \( \xi_0(y) \equiv 1 \) and \( \xi_1(y) = 1 - y \). Under the condition (A1), we may differentiate the left hand side of (2.3) with respect to \( \theta \). This yields

\[
E_{\theta_0} \xi^{(1)} \left[ K \left( \frac{x - X_1}{h} \right), E_{\theta} K \left( \frac{x - X_1}{h} \right) \right] E_{\theta} K \left( \frac{x - X_1}{h} \right) u(X_1, \theta)
\]

\[
= -E_{\theta_0} \xi \left[ K \left( \frac{x - X_1}{h} \right), E_{\theta} K \left( \frac{x - X_1}{h} \right) \right] u(X_1, \theta)
\]

(2.6)

for all \( \theta \), where \( \xi^{(1)}(y, z) = (\partial/\partial z) \xi(y, z) \).

Below, we give some preliminary results for the discussion in Section 3. Let \( \theta_h(x) \) denote the solution of the equation

\[
E \hat{\Psi}_h(x, \theta) = 0.
\]

(2.7)

Also, define a population version of \( \hat{f}_h(x) \) by \( f_h(x) = f(x, \theta_h(x)) \). These two quantities are those to which \( \hat{\theta}_h(x) \) and \( \hat{f}_h(x) \), respectively, get closer as the sample size \( n \) grows.

Next, define \( \theta_\infty \) to be the solution of

\[
Eu(X, \theta) = 0.
\]

(2.8)

Later, it will be seen that \( \theta_\infty \) is the limit of \( \theta_h(x) \) as \( h \) tends to infinity. If the true density \( g \) belongs to the parametric model, i.e., if it equals \( f(\cdot, \theta) \) for some \( \theta \), then \( \theta_\infty \) equals that value of the parameter and \( g = f(\cdot, \theta_\infty) \). Throughout the paper, we assume that the true density lies in a \( n^{-(1+\alpha)/2} \) neighborhood of the parametric model in the following sense.

(A3) The true density \( g \equiv g_n \) satisfies for some \( \alpha > -1 \) and \( c > 0 \)

\[
\int \left( g(x) - f(x, \theta_\infty) \right)^2 \, dx \leq cn^{-(1+\alpha)}.
\]

A relevant expansion of \( \xi[K(h^{-1}(x - X_1)), E_{\theta_\infty} K(h^{-1}(x - X_1))] \) plays an important role for the asymptotic analysis in Section 3. Define

\[
Y_\infty(x) = \frac{\|x - X_1\|^2}{E_{\theta_\infty} \|x - X_1\|^2}.
\]

This is the limit of \( Y_h(x) = \{1 - K(h^{-1}(x - X_1))\}/\{1 - E_{\theta_\infty} K(h^{-1}(x - X_1))\} \) as \( h \) tends to infinity. Also, let

\[
W(x) = \frac{\kappa_4}{\kappa_2} \xi_0'(Y_\infty(x)) Y_\infty(x) \left[ E_{\theta_\infty} \{Y_\infty(x) \|x - X_1\|^2\} - \|x - X_1\|^2 \right]
\]

\[
+ \kappa_2 \xi_1(Y_\infty(x)) E_{\theta_\infty} \{\|x - X_1\|^2\}
\]

(2.9)
In the following lemma we give a useful expansion for $\xi$.

**Lemma 1.** In addition to the conditions (A1) and (A2), assume that $g$ has compact support. Then, for any constant $c > 0$ there exists $\epsilon_h$ going down to zero as $h$ tends to infinity such that

$$P \left[ \sup_{\|x\| < c} \left| \xi \left( K \left( \frac{x - X_1}{h} \right), E_{\theta_\infty} \left( \frac{x - X_1}{h} \right) \right) - \xi_0(Y_\infty(x)) - \frac{1}{h^2} W(x) \right| \leq \frac{\epsilon_h}{h^2} \right] = 1.$$

**Proof.** Let $z_h(x) = 1 - E_{\theta_\infty} K(h^{-1}(x - X_1))$. Then, we can write

$$\xi \left( K \left( \frac{x - X_1}{h} \right), E_{\theta_\infty} \left( \frac{x - X_1}{h} \right) \right) = \xi \left[ 1 - Y_h(x) z_h(x), 1 - z_h(x) \right]. \quad (2.10)$$

From the condition (A2) and compactness of the support of $g$, it follows that for any constant $c > 0$ there exists $\epsilon_h$ going down to zero as $h$ tends to infinity such that with probability one

$$\sup_{\|x\| < c} \left| Y_h(x) - Y_\infty(x) \right| \leq \frac{\epsilon_h}{h^2}. \quad (2.11)$$

Applying (2.5) to (2.10) with the fact that $\sup_{\|x\| < c} |z_h(x)| \leq c'/h^2$ for some $c' > 0$, and using (2.11) yields the lemma. □

**Remark 1.** The condition that $g$ has compact support in Lemma 1 may be relaxed to a weaker one. In fact, it may be proved with more deliberate arguments that the lemma still holds when $g$ has exponentially decaying tails. However, in this case one needs some stronger conditions on $\xi$, instead. For example, in place of (2.5) one needs

$$\lim_{z \to 0} \sup_{0 \leq y \leq z^{-1+\beta}} \frac{1}{z} \left| \xi(1 - y z, 1 - z) - \xi_0(y) - \xi_1(y) z \right| = 0$$

for an arbitrarily small $\beta > 0$. In addition, to control its property at tails one needs

$$\sup_{0 < z < \varepsilon} \sup_{z^{-1+\beta} \leq \frac{y}{z} \leq z^{-1}} z^\gamma \left| \xi(1 - y z, 1 - z) - \xi_0(y) - \xi_1(y) z \right| < c$$

for some $\varepsilon, c > 0$ and $\gamma \geq 0$. The three versions of the local likelihood estimation still satisfy these conditions on $\xi$: for the U-version, $c = 1$ and $\gamma = 0$; for the C-version, $c = 1$ and $\gamma = 1$; and for the T-version, $c = (1 - \varepsilon)^{-1}$ and $\gamma = 0$. □
The following two lemmas demonstrate the behaviour of \( f_h(x) \), defined immediately below (2.7), and that of \( \hat{f}_h(x) \) when \( h \) tends to infinity. These are useful to quantify the asymptotic risk of the estimator in the next section. To state the lemmas, let

\[
U_h(x, \theta) = \hat{\Psi}_h(x, \theta) - E\hat{\Psi}_h(x, \theta).
\]

It has mean zero and variance of order \( O(n^{-1}) \). Define

\[
I_h(x, \theta) = E\left\{K\left(\frac{x-X_1}{h}\right)u(X_1, \theta)u(X_1, \theta)^T\right\} - E\left\{\frac{\hat{f}}{f}(X_1, \theta)\right\} - E\left\{\xi \left[K\left(\frac{x-X_1}{h}\right), E\theta K\left(\frac{x-X_1}{h}\right)\right] u(X_1, \theta)\right\} \times E\theta \left\{K\left(\frac{x-X_1}{h}\right)u(X_1, \theta)^T\right\}.
\]

(2.12)

It will be seen that this is an approximation of \(-\frac{\partial}{\partial \theta}E\hat{\Psi}_h(x, \theta)\). Above and in the subsequent arguments, we let \( \dot{p}(x, \theta) \) and \( \ddot{p}(x, \theta) \) for a function \( p \) denote, respectively, the first and the second derivatives of \( p(x, \theta) \) with respect to \( \theta \). Write \( f_\infty(x) = f(x, \theta_\infty) \), and define a \( p \times p \) matrix

\[
I_\infty(\theta) = E\left\{u(X_1, \theta)u(X_1, \theta)^T - \frac{\hat{f}}{f}(X_1, \theta)\right\}.
\]

The two lemmas rely on some technical assumptions in addition to (A1)-(A3). They are stated in the appendix. Proofs of the lemmas are also deferred to the appendix.

**Lemma 2.** Assume (A1)-(A3) and the conditions listed in the appendix. Then, for all \( \alpha \geq -1 \) it follows that uniformly for \( x \) in any compact subset of \( \mathcal{X} \)

\[
f_h(x) = f_\infty(x) + \dot{f}(x, \theta_\infty)^T I_\infty(\theta_\infty)^{-1} E\hat{\Psi}_h(x, \theta_\infty) + O(\rho) \tag{2.13}
\]

as \( n \to \infty \) and \( h \to \infty \), where \( \rho \equiv \rho(n, h) = n^{-1+\alpha/2}h^{-4} \).

**Lemma 3.** Assume the conditions of Lemma 2. Then, for all \( \alpha \geq -1 \) it follows that uniformly for \( x \) in any compact subset of \( \mathcal{X} \)

\[
\hat{f}_h(x) = f_h(x) + \dot{f}(x, \theta_\infty)^T I_h(x, \theta_\infty)^{-1} U_h(x, \theta_\infty) + O_p(\delta)
\]

as \( n \to \infty \) and \( h \to \infty \), where \( \delta \equiv \delta(n, h) = n^{-1-\alpha/2}h^{-2}(\log n)^{1/2} + n^{-1} \log n \).

**Remark 2.** Since \( U_h(x, \theta_\infty) \) at Lemma 3 is a sum of independent random vectors, asymptotic normality of \( \hat{f}_h(x) \) follows immediately from the lemma.

3. Risk analysis
For the risk of an estimator \( \hat{g} \) of \( g \), we consider the mean integrated weighted squared error:

\[
E \int \{ \hat{g}(x) - g(x) \}^2 w(x) \, dx,
\]

where \( w(\cdot) \) is a weight function whose support is compact and contained in the support of \( g \). In this section, we provide the asymptotic risks of the estimator \( \hat{f}_h(x) \) and its scaled version \( \hat{g}_h(x) \).

First, we consider the unscaled estimator \( \hat{f}_h(x) \). The asymptotic risk of the estimator \( \hat{f}_h(x) \) is given by the risk of its approximation \( \tilde{f}_h(x) \) which is defined by

\[
\tilde{f}_h(x) = f_h(x) + \hat{f}(x, \theta_\infty)^T I_h(x, \theta_\infty)^{-1} U_h(x, \theta_\infty).
\]

(3.1)

We decompose the risk of \( \tilde{f}_h \) into two parts:

\[
\begin{align*}
b_U(n, h) &= \int \{ f_h(x) - g(x) \}^2 w(x) \, dx, \\
v_U(n, h) &= E \int \{ \tilde{f}_h(x) - f_h(x) \}^2 w(x) \, dx.
\end{align*}
\]

The first term \( b_U(n, h) \) represents the bias of the unscaled estimator \( \tilde{f}_h \) due to model misspecification, while \( v_U(n, h) \) measures its sampling variability.

In the following theorem, we give approximations for these components. To state the theorem, let \( I_{0, \infty} = Eu(X_1, \theta_\infty)u(X_1, \theta_\infty)^T \) and \( D_\infty = -E\{ \hat{f}(X_1, \theta_\infty)/f(X_1, \theta_\infty) \} \). Note that \( D_\infty \) is not zero since we take the expectation with respect to the true density instead of \( f(\cdot, \theta_\infty) \). It follows that

\[
I_{0, \infty} = I_\infty(\theta_\infty) - D_\infty.
\]

(3.2)

Define

\[
\begin{align*}
\nu_0 &= \int \{ f_{\infty}(x) - g(x) \}^2 w(x) \, dx, \\
\tau_0 &= \int \hat{f}(x, \theta_\infty)^T I_{0, \infty}(\theta_\infty)^{-1} I_{0, \infty}(\theta_\infty)^{-1} \hat{f}(x, \theta_\infty) w(x) \, dx, \\
U(x) &= \hat{f}(x, \theta_\infty)^T I_{0, \infty}^{-1} D_\infty I_{0, \infty}^{-1} u(X_1, \theta_\infty), \\
Z_U(x) &= \hat{f}(x, \theta_\infty)^T I_{0, \infty}^{-1} \left\{ \| x - X_{1, \infty} \|^2 u(X_1, \theta_\infty) - \xi_0(Y_{\infty}(x)) E\theta_\infty \| x - X_1 \|^2 u(X_1, \theta_\infty) \right\}, \\
\nu_{1, U} &= 2 \int \left\{ E_{g-f_{\infty}} Z_U(x) \right\} \{ f_{\infty}(x) - g(x) \} w(x) \, dx, \\
\tau_{1, U} &= 2 \int \left\{ E Z_U(x) U(x) \right\} w(x) \, dx, \\
\tau_{2, U} &= \int \left[ E Z_U(x)^2 - \left\{ E Z_U(x) U(X_1, \theta_\infty)^T \right\} I_{0, \infty}^{-1} \left\{ E u(X_1, \theta_\infty) Z_U(x) \right\} \right] w(x) \, dx.
\end{align*}
\]
Here \( E_{g-f_\infty} \) denotes \( E - E_{\theta_\infty} \). Note that \( \nu_0 = O(n^{-1-\alpha}) \) and \( \nu_{1, U} = O(n^{-1-\alpha}) \) by the condition (A3). Also, \( \tau_{1, U} = O(n^{-(1+\alpha)/2}) \) since \( D_\infty = O(n^{-(1+\alpha)/2}) \) by the fact \( E_{\theta_\infty} \{ \hat{f}(X_1, \theta_\infty) / f(X_1, \theta_\infty) \} = 0 \). Now, \( \tau_{2, U} \) is a constant which does not depend on \( n \). It is strictly positive. This follows from the Cauchy-Schwartz inequality: for any \( \mathbb{R}^p \)-valued \( \psi \) and real-valued \( \phi \)
\[
\left( E\phi \psi^T \right) \left( E\psi \psi^T \right)^{-1} \left( E\psi \phi \right) \leq E\phi^2
\]
with ‘=’ holding if and only if \( \phi = a^T \psi \) for some constant vector \( a \). Taking \( \psi = U(X_1, \theta_\infty) \) and \( \phi = \hat{U}_\nu(x) \) shows that \( \tau_{2, U} > 0 \).

**Theorem 1.** Under the conditions of Lemma 2, we get as \( n \to \infty \) and \( h \to \infty \)
\[
\begin{align*}
b_U(n, h) &= \nu_0 - \kappa_2 \frac{\nu_{1, U}}{h^2} + o \left( \frac{1}{n^{\alpha+1}h^2} \right), \\
v_U(n, h) &= \frac{\tau_0}{n} - \kappa_2 \frac{\tau_{1, U}}{nh^2} + \frac{\kappa_2 \tau_{2, U}}{nh^4} + o \left( \frac{1}{n^{3/2+\alpha/2}h^2} + \frac{1}{nh^4} \right).
\end{align*}
\]

Next, we consider the scaled estimator \( \hat{g}_h(x) \). Define \( g_h(x) = f_h(x) / \int f_h(x) \, dx \).
Write
\[
\begin{align*}
v_h(x) &= \hat{f}(x, \theta_\infty)^T I_\infty(\theta_\infty)^{-1} E\hat{U}_\nu(x, \theta_\infty), \\
V_h(x) &= \hat{f}(x, \theta_\infty)^T I_h(x, \infty)^{-1} U_h(x, \theta_\infty).
\end{align*}
\]
We may obtain expansions for \( g_h(x) \) and \( \hat{g}_h(x) \), analogous to those given at Lemmas 2 and 3, as follows:
\[
\begin{align*}
g_h(x) &= f_\infty(x) + v_h(x) - f_\infty(x) \int v_h(x) \, dx + O(\rho), \quad (3.3) \\
\hat{g}_h(x) &= g_h(x) + V_h(x) - f_\infty(x) \int V_h(x) \, dx + O_p(\delta). \quad (3.4)
\end{align*}
\]
In the next theorem, we give the risk of \( \tilde{g}_h(x) \), an approximation of \( \hat{g}_h(x) \) defined by
\[
\tilde{g}_h(x) = g_h(x) + V_h(x) - f_\infty(x) \int V_h(x) \, dx.
\]
Similarly to the case of the unscaled \( \tilde{f}_h(x) \), the risk of \( \tilde{g}_h(x) \) is decomposed into two parts:
\[
\begin{align*}
b_S(n, h) &= \int \{g_h(x) - g(x)\}^2 w(x) \, dx, \quad (3.5) \\
v_S(n, h) &= E \int \{\tilde{g}_h(x) - g_h(x)\}^2 w(x) \, dx. \quad (3.6)
\end{align*}
\]
To state the theorem, write

\[ Z_S(x) = Z_U(x) - f_\infty(x) \int Z_U(x) \, dx. \]

Define the following scaled versions of \( \nu_{1,U}, \tau_{1,U} \) and \( \tau_{2,U} \):

\[
\begin{align*}
\nu_{1,S} &= 2 \int \{ E_{g-f_\infty} Z_S(x) \} \{ f_\infty(x) - g(x) \} \, w(x) \, dx, \\
\tau_{1,S} &= 2 \int \{ E Z_S(x) U(x) \} \, w(x) \, dx, \\
\tau_{2,S} &= \int \left[ E Z_S(x)^2 - \{ E Z_S(x) u(X_1, \theta_\infty)^T \} I_{0,\infty}^{-1} \{ E u(X_1, \theta_\infty) Z_S(x) \} \right] \, w(x) \, dx.
\end{align*}
\]

**Theorem 2.** Under the conditions of Lemma 2, we get as \( n \to \infty \) and \( h \to \infty \)

\[
\begin{align*}
b_S(n, h) &= \nu_0 - \kappa_2 \frac{\nu_{1,S}}{h^2} + o \left( \frac{1}{n^{\alpha+1}h^2} \right), \\
v_S(n, h) &= \frac{\tau_0}{n} - \kappa_2 \frac{\tau_{1,S}}{nh^2} + \kappa_2 \frac{\tau_{2,S}}{nh^4} + o \left( \frac{1}{n^{3/2+\alpha/2}h^2} + \frac{1}{nh^4} \right).
\end{align*}
\]

The first term \( \nu_0 \) in the expansions of \( b_U(n, h) \) and \( b_S(n, h) \) in Theorems 1 and 2 represents the model misspecification error of the proposed parametric family. It is the integrated squared bias of the parametric maximum likelihood estimator \( f(\cdot, \hat{\theta}_{\text{MLE}}) \), where \( \hat{\theta}_{\text{MLE}} \) is defined as the solution of the equation \( \sum_{i=1}^n u(X_i, \theta) = 0 \). It is zero if the true density actually belongs to the parametric model. Next, the first term \( \tau_0/n \) in the expansions of \( v_U(n, h) \) and \( v_S(n, h) \) is the integrated variance of the parametric maximum likelihood estimator. Since \( \nu_0 \) and \( \tau_0 \) do not depend on \( \xi \), the first order properties of all the members in the class of the local likelihood estimation are the same. The other terms in the expansions depend on the bandwidth. These terms also depend on \( \xi \), but only through \( \xi_0 \). Thus, the U- and T-versions have the same second order properties, too, as they both have \( \xi_0 \equiv 1 \). We note that Eguchi and Copas (1998) neglected the term \( \kappa_2 \tau_1/nh^2 \) in their analysis of the asymptotic variance because their main concern was the case where \( 0 < \alpha < 1 \). When \( 0 < \alpha < 1 \), the optimal \( h \) is of order \( n^{\alpha/2} \) (see Section 4) and with this choice the term \( \kappa_2 \tau_1/nh^2 \) is negligible.

We may find an optimal size of the bandwidth by minimizing the sum of \( b(n, h) \) and \( v(n, h) \). This will be discussed in Section 4. The formulas given in Theorems 1 and 2 are more useful than the one given by Eguchi and Copas (1998) since dependence of the risks on the function \( \xi \) are more transparent. A direct risk comparison between various local
likelihood methods can be made from the formulas, which will be dealt too in the next section. Below, we give a proof of Theorem 1. Proof of Theorem 2 is omitted as it may be proved in a similar fashion using (3.3) and (3.4) instead of Lemmas 2 and 3.

**Proof of Theorem 1.** From the consistency condition (2.3) on \( \xi \),

\[
E \hat{\Psi}_h(x, \theta_\infty) = E_{g_{f_\infty}} \left\{ K \left( \frac{x-X_1}{h} \right) u(X_1, \theta_\infty) - \xi \left[ K \left( \frac{x-X_1}{h} \right) , E_{\theta_\infty} K \left( \frac{x-X_1}{h} \right) \right] \right\}.
\]

By similar arguments as in deriving (3.7), we get

\[
The first part of the theorem then follows immediately from (2.13) at Lemma 2 and (3.7).

Thus, from Lemma 1 and the fact \( E_{g_{f_\infty}} u(X_1, \theta_\infty) = 0 \), we obtain from the condition (A2) on the kernel that uniformly for \( x \) in any compact subset of \( X \)

\[
E_{\theta_\infty} K \left( \frac{x-X_1}{h} \right) u(X_1, \theta_\infty) = -\frac{\kappa_2}{h^2} E_{\theta_\infty} \| x - X_1 \|^2 u(X_1, \theta_\infty) + O \left( \frac{1}{h^4} \right).
\]

Thus, from Lemma 1 and the fact \( E_{g_{f_\infty}} u(X_1, \theta_\infty) = 0 \), it follows that uniformly for \( x \) in any compact subset of \( X \)

\[
\hat{J}(x, \theta_\infty)^T I_\infty(\theta_\infty)^{-1} E \hat{\Psi}_h(x, \theta_\infty) = -\frac{\kappa_2}{h^2} E_{g_{f_\infty}} Z u(x) + o \left( \frac{1}{n(1+\alpha)/2h^2} \right).
\] (3.7)

The first part of the theorem then follows immediately from (2.13) at Lemma 2 and (3.7).

To find the formula for \( v(n, h) \), we need to approximate \( I_h(x, \theta_\infty) \), defined at (2.12), and \( \text{var} \{ U_h(x, \theta_\infty) \} \). Define

\[
I_{k,\infty}(x) = E \left\{ \| x - X_1 \|^k u(X_1, \theta_\infty) u(X_1, \theta_\infty)^T \right\},
\]

\[
Z(x) = \| x - X_1 \|^2 u(X_1, \theta_\infty) - \xi_0(Y_\infty(x)) E_{\theta_\infty} \| x - X_1 \|^2 u(X_1, \theta_\infty).
\]

By similar arguments as in deriving (3.7), we get

\[
I_h(x, \theta_\infty) = I_\infty(\theta_\infty) - \frac{\kappa_2}{h^2} E Z(x) u(X_1, \theta_\infty)^T + \frac{1}{h^4} \left[ \kappa_4 I_{4,\infty}(x) \right]
+ \kappa_2 EW(x) u(X_1, \theta_\infty) \left\{ E_{\theta_\infty} \| x - X_1 \|^2 u(X_1, \theta_\infty)^T \right\} \] (3.8)

uniformly for \( x \) in any compact subset of \( X \).

We compute \( \text{var} \{ U_h(x, \theta_\infty) \} \). From Lemma 1 and the condition (A2) on the kernel \( K \), we may verify that uniformly for \( x \) in any compact subset of \( X \)

\[
\text{var} \{ U_h(x, \theta_\infty) \} = \frac{1}{n} \text{var} \left\{ u(X_1, \theta_\infty) - \frac{\kappa_2}{h^2} Z(x) + \frac{1}{h^4} \left[ \kappa_4 \| x - X_1 \|^4 u(X_1, \theta_\infty) \right. \right.
+ \kappa_2 W(x) E_{\theta_\infty} \| x - X_1 \|^2 u(X_1, \theta_\infty) \left\} \right\} + o \left( \frac{1}{nh^4} \right). \] (3.9)
Next, write $H_2(x) = \kappa_2 E Z(x) u(X_1, \theta_\infty)^T$ and

$$H_4(x) = \kappa_4 I_{4,\infty}(x) + \kappa_2 E W(x) u(X_1, \theta_\infty) \left\{ E_{\theta_\infty} \| x - X_1 \|^2 u(X_1, \theta_\infty)^T \right\}.$$

Note that with these notations the equation (3.8) can be written as

$$I_h(x, \theta_\infty) = I_\infty(\theta_\infty) - \frac{1}{h^2} H_2(x) + \frac{1}{h^4} H_4(x). \tag{3.10}$$

Also, the equation (3.9) reduces to

$$n \text{ var } \{U_h(x, \theta_\infty)\} = I_{0,\infty} - \frac{1}{h^2} \left\{ H_2(x) + H_2(x)^T \right\} + \frac{1}{h^4} \left\{ H_4(x) + H_4(x)^T \right\}
+ \frac{\kappa_2^2}{h^4} E Z(x) Z(x)^T \right\} + o \left( \frac{1}{h^4} \right). \tag{3.11}$$

From (3.10) it follows that uniformly for $x$ in any compact subset of $\mathcal{X}$

$$I_h(x, \theta_\infty)^{-1} = I_\infty(\theta_\infty)^{-1} + \frac{1}{h^2} I_\infty(\theta_\infty)^{-1} H_2(x) I_\infty(\theta_\infty)^{-1}
+ \frac{1}{h^4} I_\infty(\theta_\infty)^{-1} \left\{ H_2(x) I_\infty(\theta_\infty)^{-1} H_2(x) - H_4(x) \right\} I_\infty(\theta_\infty)^{-1} \tag{3.12}$$

$$+ o \left( \frac{1}{h^4} \right).$$

We plug (3.11) and (3.12) into

$$n \tilde{f}(x, \theta_\infty)^T I_h(x, \theta_\infty)^{-1} \text{ var } \{U_h(x, \theta_\infty)\} \{I_h(x, \theta_\infty)^{-1}\}^T \tilde{f}(x, \theta_\infty),$$

and collect terms involving $h^{-2}$ and $h^{-4}$. We find that the $h^{-2}$ terms are

$$\frac{1}{h^2} \tilde{f}(x, \theta_\infty)^T I_\infty(\theta_\infty)^{-1} \left\{ H_2(x) I_\infty(\theta_\infty)^{-1} I_{0,\infty} + I_{0,\infty} I_\infty(\theta_\infty)^{-1} H_2(x)^T
- H_2(x) - H_2(x)^T \right\} I_\infty(\theta_\infty)^{-1} \tilde{f}(x, \theta_\infty) \tag{3.13}$$

$$= \frac{2}{h^2} \tilde{f}(x, \theta_\infty)^T I_\infty(\theta_\infty)^{-1} \left\{ H_2(x) I_\infty(\theta_\infty)^{-1} E \frac{\tilde{f}(X_1, \theta_\infty)}{\tilde{f}(x, \theta_\infty)} \right\} I_\infty(\theta_\infty)^{-1} \tilde{f}(x, \theta_\infty).$$

The equation (3.13) follows from (3.2). We can replace $I_\infty(\theta_\infty)^{-1}$ in (3.13) by $I_{0,\infty}^{-1}$ with an error $O(n^{-1-\alpha} h^{-2})$ since $E\{\tilde{f}(X_1, \theta_\infty)/\tilde{f}(X_1, \theta_\infty)\} = O(n^{-(1+\alpha)/2})$. This gives that (3.13) equals $-2 \kappa_2 h^{-2} E\{Z_U(x)u(x)\} + O(n^{-1-\alpha} h^{-2})$. Similarly, we find that the $h^{-4}$ terms reduce to

$$\frac{1}{h^4} \tilde{f}(x, \theta_\infty)^T I_\infty(\theta_\infty)^{-1} \left[ \kappa_2^2 E Z(x) Z(x)^T - H_2(x) I_\infty(\theta_\infty)^{-1} H_2(x)^T \right]
\times I_\infty(\theta_\infty)^{-1} \tilde{f}(x, \theta_\infty) + o \left( \frac{1}{h^4} \right)$$

$$= \frac{\kappa_2^2}{h^4} \left[ E Z_U(x)^2 - \{E Z_U(x) u(X_1, \theta_\infty)^T \} I_{0,\infty}^{-1} \{E u(X_1, \theta_\infty) Z_U(x)\} \right] + o \left( \frac{1}{h^4} \right).$$
uniformly for $x$ in any compact subset of $\mathcal{X}$. The second part of the theorem now follows.

**Remark 3 (Kullback-Leibler risk).** An expansion of the Kullback-Leibler risk $E \int \log \{g(x)/\hat{g}_h(x)\} g(x) \, dx$ may be derived from Theorem 2 with specific choices of the weight function $w$. In fact, under the condition that $f_\infty$ is bounded away from zero on the support of $g$, we may approximate the Kullback-Leibler risk by

$$\text{KL}(g, f_\infty) - \int \left\{ \frac{g_h(x) - f_\infty(x)}{f_\infty(x)} \right\} g(x) \, dx + \frac{1}{2} E \int \left\{ \frac{\hat{g}_h(x) - g_h(x)}{g_h(x)} \right\}^2 g(x) \, dx$$

(3.14)

where $\text{KL}(g, f) = \int \log \{g(x)/f(x)\} g(x) \, dx$ denotes the Kullback-Leibler divergence of $f$ from $g$. The first term $\nu_0^{\text{KL}} \equiv \text{KL}(g, f_\infty)$ at (3.14) is the minimal Kullback-Leibler divergence from $g$ among all members in $\{f(\cdot, \theta) : \theta \in \Theta\}$. The second term equals

$$- \int \{g_h(x) - f_\infty(x)\} \{g(x) - f_\infty(x)\} f_\infty(x)^{-1} \, dx$$

(3.15)

since $\int \{g_h(x) - f_\infty(x)\} \, dx = 1 - 1 = 0$. From the definition of $b_S(n,h)$ at (3.5) and its expansion given at Theorem 2, we may get an approximation of (3.15). By applying the first part of Theorem 2 with $w(x) = 1/\{2f_\infty(x)\}$, we obtain (3.15) equals $-\kappa_2 \nu_1^{\text{KL}} h^{-2} + o(n^{-(1+\alpha)}h^{-2})$ where

$$\nu_1^{\text{KL}} = \int \{E_{g-f_\infty} Z_S(x)\} \{f_\infty(x) - g(x)\} f_\infty(x)^{-1} \, dx.$$

When $\xi_0 \equiv 1$ (thus for U- and T-version), it may be proved that

$$\nu_1^{\text{KL}} = 2 \{E_{g-f_\infty} X_1 u(X_1, \theta_\infty)\}^T I_{0,\infty}^{-1} \{E_{g-f_\infty} X_1 u(X_1, \theta_\infty)\}. \quad (3.16)$$

This matches the bias results in Theorem 1 and Corollary 1 of Eguchi and Copas (1998). This means that the results of Eguchi and Copas (1998) are valid only for the case where $\xi_0 \equiv 1$. Note that the unscaled version $f_h$ does not integrate to one. Thus, the second term at (3.14) corresponding to the unscaled estimator $\hat{f}_h$ would be of order $n^{-(1+\alpha)/2}h^{-2}$ which is slower than $n^{-(1+\alpha)}h^{-2}$ of the scaled estimator. Therefore, normalizing the local likelihood estimator by its integral is important for the Kullback-Leibler risk. The third term at (3.14) is the variance part. We may get an expansion of this from the second part of Theorem 2, now with $w(x) = g(x)/\{2f_\infty(x)^2\}$. It equals

$$\frac{\tau_0^{\text{KL}}}{n} - \kappa_2 \tau_1^{\text{KL}} \frac{1}{nh^2} + \kappa_2 \tau_2^{\text{KL}} \frac{1}{nh^4} + o \left( \frac{1}{n^{3/2+\alpha/2}h^2} + \frac{1}{nh^4} \right)$$
where $\tau_i^{KL}$ for $i = 0, 1, 2$ are defined as $\tau_0$, $\tau_1,S$ and $\tau_2,S$ with $g(x)/\{2f_\infty(x)^2\}$ replacing $w(x)$. For instance,

$$\tau_0^{KL} = \frac{1}{2} \int u(x, \theta_\infty)^T I_\infty(\theta_\infty)^{-1} I_{0,\infty}(\theta_\infty)^{-1} u(x, \theta_\infty) g(x) \, dx. \quad \blacksquare$$

4. Comparison and optimal bandwidth

In Theorems 1 and 2, $\nu_0 + (\tau_0/n)$ is the mean integrated squared error of the parametric maximum likelihood estimator. Thus, the risk improvement achieved by the local likelihood estimators upon the parametric maximum likelihood estimator is given by

$$r_d(h) \equiv \kappa_2^{2} (\nu_1 + \tau_1 n^{-1}) h^{-2} - \kappa_2^{2} \tau_2 n^{-1} h^{-4}. \quad (4.1)$$

Here and below in this section, we simply write $\nu_i$ and $\tau_i$ ($i = 1, 2$) for $\nu_{1,k}$ and $\tau_{i,k}$, respectively, where $k = U$ or $S$. As a function of $t = h^{-2}$, $r_d(h)$ is a concave parabola on $t \geq 0$. It has the maximum value at $t_0 = (\nu_1 n + \tau_1)/(2\kappa_2 \tau_2)$ if $\nu_1 n + \tau_1 > 0$. Thus, in this case the optimal bandwidth is given by

$$h_{opt} = \left( \frac{2\kappa_2 \tau_2}{\nu_1 n + \tau_1} \right)^{1/2}, \quad (4.2)$$

and the maximum risk improvement equals $(\nu_1 n + \tau_1)^2/(4n\tau_2)$. When $\nu_1 n + \tau_1 \leq 0$, the risk improvement $r_d$ is a strictly decreasing function of $t$ on $t \geq 0$, thus it is maximized at $t = 0$, i.e. at $h = \infty$ with the maximum value being zero. Note that $h = \infty$ corresponds to the fully parametric maximum likelihood estimator.

It is not clear to us whether $\nu_1 n + \tau_1 > 0$ in general. However, we found in an example below $\nu_1$ and $\tau_1$ are positive (see Figure 2). For the Kullback-Leibler risk, it can be seen from (3.16) that the U- and T-versions have $\nu_1^{KL} + \tau_1^{KL} > 0$ for sufficiently large $n$. In the subsequent discussion we assume $\nu_1 n + \tau_1 > 0$. Now, recall that $\nu_1 \asymp n^{-1+\alpha}$ and $\tau_1 \asymp n^{-(1+\alpha)/2}$. Thus, the formula (4.2) is valid only when $\alpha > 0$ since it is derived in the large $h$ setting where $h$ tends to infinity as $n$ grows. If $0 < \alpha < 1$, then $(\nu_1 n)$ dominates $\tau_1$. Thus, in this case $h_{opt} \asymp n^{\alpha/2}$ and the maximum risk improvement is of order $n^{-(1+2\alpha)}$.

Next, when $\alpha \geq 1$, the optimal bandwidth is asymptotic to $n^{(1+\alpha)/4}$ with the maximum risk improvement being of order $n^{-(2+\alpha)}$. In the remaining case where $-1 < \alpha \leq 0$, we see from (4.1) that letting $h$ tend to infinity at a slower rate makes $r_d$ larger. Thus, a bandwidth tending to infinity at an ultimately slow rate would be preferable in this case.
We combine the results of Park et al. (2002) into our large \( h \) analysis. Recall \( d \) is the dimension of \( X_i \) and \( p \) is the dimension of the parameter. When \( d = 1 \), it was shown that the optimal bandwidth in the small \( h \) setting is asymptotic to \( n^{-1/\{1+4[(p+1)/2]\}} \) with the minimal risk being of order \( n^{-q_1} \), where \( q_1 = 4[(p+1)/2]/\{1+4[(p+1)/2]\} \) and \( [(p+1)/2] \) denotes the greatest integer which is less than or equal to \( (p+1)/2 \). This can be generalized to an arbitrary \( d \). Let \( q_d = 4[(p+1)/2]/\{d+4[(p+1)/2]\} \). It can be seen that in the \( d \)-variate case the minimum risk \( n^{-q_d} \) is achieved by the optimal bandwidth of order \( n^{-1/\{d+4[(p+1)/2]\}} \). Note that the first order in the risk expansion for large \( h \) is \( n^{-(1+\alpha)} + n^{-1} \). Comparing this with the small \( h \) optimal risk \( n^{-q_d} \) and taking into account the discussion in the previous paragraph, we arrive at the following conclusion. We find that the value of \( \alpha \) at which a transition from a small to a large bandwidth is desirable is \( \alpha = q_d - 1 \).

(i) \(-1 \leq \alpha < q_d - 1\): \( h_{opt} \asymp n^{-1/\{d+4[(p+1)/2]\}} \);
(ii) \( q_d - 1 < \alpha \leq 0\): \( h \) tending to infinity at an ultimately slow rate is preferable;
(iii) \( 0 < \alpha < 1\): \( h_{opt} \asymp n^{\alpha/2} \);
(iv) \( 1 \leq \alpha\): \( h_{opt} \asymp n^{(1+\alpha)/4} \).

The large \( h \) asymptotic formula (4.2) and the small \( h \) results provided in Park et al. (2002) may be used to produce useful bandwidth selectors. For example, plug-in methods are immediate from the formula (4.2), where \( \theta_\infty \) is replaced by the solution of the likelihood equation \( \sum_{i=1}^{n} u(X_i, \theta) = 0 \) and other unknown quantities by their obvious empirical versions. Least squares cross-validation is an alternative way of choosing a data-driven bandwidth selector, and is readily applicable to local likelihood density estimation. The latter is not so tied to asymptotics and does not depend on the knowledge of \( \alpha \). Thus, it may be used for a goodness-of-fit test of a parametric model, where the parametric model is rejected for small values of cross-validatory bandwidth selector. Determination of the cut-off values in this case requires the sampling distribution of the bandwidth selector. This would be a challenging problem for future research.

5. A skewed normal example

We compare the large \( h \) properties of the U- and C-versions of the local likelihood estimation. Note that the T-version has the same first and second order properties with
the U-version as we pointed out in the paragraph immediately after the statement of Theorem 2. We consider \( N(\theta, 1) \) as the parametric model. We take \( w(x) \equiv 1 \) in the definition of the risk. The true density is taken to be

\[
g(x) \equiv g_\beta(x) = 2\phi(x)\Phi(\beta x),
\]

(4.1)

where \( \phi \) and \( \Phi \) are the standard normal density and its distribution function. This is the so-called skewed normal distribution of Azzalini (1985), and was also considered by Eguchi and Copas (1998). Here, \( \beta \) acts as a discrepancy parameter. When \( \beta = 0 \), the density \( g \) is identical to \( \phi \). As \( |\beta| \) increases, it becomes increasingly skewed. In this setting, we find

\[
\theta_\infty = EX = \sqrt{\frac{2}{\pi}} \frac{\beta}{\sqrt{1 + \beta^2}}.
\]

(4.2)

The integrated squared distance between the true density \( g \) and its best parametric approximant \( \phi(\cdot - \theta_\infty) \), which is \( \nu_0 = \int \{ \phi(x - \theta_\infty) - 2\phi(x)\Phi(\beta x) \}^2 \, dx \), is a symmetric function of \( \beta \). Figure 1 depicts \( \nu_0 \) as a function of \( \beta \).

We calculate some ingredients to evaluate the risks given in Theorems 1 and 2. We find

\[
I_\infty(\theta_\infty) = 1, I_{0,\infty} = 1 - \theta^2_\infty \text{ and } D_\infty = \theta^2_\infty.
\]

For computing \( \nu_i \) and \( \tau_i \), we use the formula for the odd moments of the skewed normal distribution given in Corollary 4 of Henze (1986). In particular, we find in addition to (4.2)

\[
EX^3 = \sqrt{\frac{2}{\pi}} \left\{ - \left( \frac{\beta}{\sqrt{1 + \beta^2}} \right)^3 + 3 \left( \frac{\beta}{\sqrt{1 + \beta^2}} \right) \right\}
\]

\[
EX^5 = \sqrt{\frac{2}{\pi}} \left\{ 3 \left( \frac{\beta}{\sqrt{1 + \beta^2}} \right)^5 - 10 \left( \frac{\beta}{\sqrt{1 + \beta^2}} \right)^3 + 15 \left( \frac{\beta}{\sqrt{1 + \beta^2}} \right) \right\}.
\]

For the even moments we obtain \( EX^{2k} = (2k)!/(2^k k!) \), and thus \( EX^2 = 1, EX^4 = 3, EX^6 = 15 \). These formula may be also obtained by applying Corollaries 3.2 and 5.3 of Aldershof et al. (1995).

It may be seen that all \( \nu_i \) and \( \tau_i \) are symmetric as functions of \( \beta \). Furthermore,

\[
E \int Z_U(y) \, dy = \int \phi(x - \theta_\infty) E(X_1 - \theta_\infty)^3,
\]

where \( C = 1 \) for the U-version and \( C = \int \phi(y^2/(1 + y^2)) \, dy \) for the C-version. Thus, since \( \tau_{1,S} = \tau_{1,U} - 2 \int f_\infty(x) \{ E \int Z_U(y) \, dy \} \, dx \) and \( f_\infty(x) = \phi(x - \theta_\infty) \), we have
\(\tau_{1,S} = \tau_{1,U}\) for both the U- and C-versions. Similarly, we may find \(\tau_{2,S} = \tau_{2,U}\), but in this case only for the U-version. Figure 2 shows \(\nu_1, \tau_1\) and \(\tau_2\). We find that \(\nu_1\) and \(\tau_1\) are positive and converge to zero as \(\beta\) tends to zero. Also, from Figure 2(a) we find that the U-versions have less bias than the C-version, and that the scaling improves the bias. If one plugs the values of \(\nu_1, \tau_1\) and \(\tau_2\) into the formula (4.2), one may see how \(h_{opt}\) changes as \(\beta\) increases. We found that for a sample of size 100 the optimal bandwidth for the U-version of \(\hat{f}_h\) decreases from infinity to .31 as \(\beta\) increases from zero to 10, and for the C-version it takes values from infinity to .30.

(Insert Figure 2 about here)

Now, we evaluate the maximum risk improvements \(r_d(h_{opt}) = (\nu_1 n + \tau_1)^2/(4n\tau_2)\) achieved by the U- and C-versions upon the parametric maximum likelihood estimator. Note that the maximum risk improvement does not depend on the choice of kernel. It is symmetric about zero as a function of \(\beta\). Figure 3 depicts \(r_d(h_{opt})\) when \(n = 100\) and 400. Comparing the scaled estimator \(\hat{g}_h\) with the unscaled \(\hat{f}_h\), we find both U(T)- and C-versions of \(\hat{g}_h\) outperform the corresponding versions of \(\hat{f}_h\) for all \(\beta\). Also, it is interesting to find that the U(T)-version is better than the C-version for all \(\beta\) in the case of \(\hat{g}_h\), but that the risks of the unscaled estimators \(\hat{f}_h\) are indistinguishable although the C-version now is slightly better. We found that this is true for other sample sizes, too.

(Insert Figure 3 about here)

We conducted a small simulation to check the validity of our discussion on the optimal bandwidth size. For this, we took the standard normal density as the kernel function. We calculated an optimal bandwidth which minimizes the sum of squared deviations \(\sum_{i=1}^{n}(\hat{f}_h(X_i) - g_\beta(X_i))^2\). Our simulation consists of the three steps; (i) for each \(\beta = 0, 0.5, 1, 2, 5,\) and 10, generate a random sample of size \(n = 100\) from the density \(g_\beta\) at (4.1) by the rejection method; (ii) compute \(\hat{f}_h\) for the U- and C-versions at each data point \(X_1, X_2, \ldots, X_n\); (iii) find the optimal bandwidth over the interval \((0, 30)\) which minimizes the sum of squared deviations \(\sum_{i=1}^{n}(\hat{f}_h(X_i) - g_\beta(X_i))^2\). These steps were repeated 100 times. Table 1 shows the average of the 100 calculated optimal bandwidths for each value of \(\beta\). We see that it clearly justifies our theoretical observation that the optimal bandwidth traverses from a large to a small value as the degree of discrepancy from the
parametric model (in this example, $\beta$) increases. We note that the theoretical values .31 and .30 at $\beta = 10$ discussed two paragraphs above do not match well with .13 in the table because the theoretical values are obtained from the formula that is valid in the near parametric case.

(Insert Table 1 about here)

Table 1: Average of the optimal bandwidth

<table>
<thead>
<tr>
<th>$\beta$</th>
<th>0.0</th>
<th>0.5</th>
<th>1.0</th>
<th>2.0</th>
<th>5.0</th>
<th>10.0</th>
</tr>
</thead>
<tbody>
<tr>
<td>$h$</td>
<td>14.15</td>
<td>12.3</td>
<td>3.08</td>
<td>0.52</td>
<td>0.16</td>
<td>0.13</td>
</tr>
</tbody>
</table>

C-version

<table>
<thead>
<tr>
<th>$\beta$</th>
<th>0.0</th>
<th>0.5</th>
<th>1.0</th>
<th>2.0</th>
<th>5.0</th>
<th>10.0</th>
</tr>
</thead>
<tbody>
<tr>
<td>$h$</td>
<td>16.25</td>
<td>13.56</td>
<td>4.55</td>
<td>0.2</td>
<td>0.15</td>
<td>0.13</td>
</tr>
</tbody>
</table>
Appendix

A.1. Additional assumptions. In addition to the assumptions (A1)–(A3), we need the following assumptions for the lemmas and the theorem.

(A4) the solution $\theta_h(x)$ defined at (2.7) is unique;

(A5) the underlying density $g$ and its best parametric approximation $f_\infty$ have compact supports;

(A6) $f(x, \theta)$ is three times partially differentiable with respect to $\theta$ and all the partial derivatives are continuous in $x$ and $\theta$;

(A7) there exists a function $G$ which is continuous and satisfies for all $x$

$$
\sup_{\theta \in \Theta} \left| \frac{\partial^2}{\partial \theta^2} u(x, \theta) \right| \leq G(x);
$$

A.2. Proof of Lemma 2. Here, all ‘$O$’ expressions are uniform for $x$ in any compact subset $S$ of $X$, i.e. for a sequence of functions $Q_{n,h}$ we say simply $Q_{n,h}(x) = O\{r(n,h)\}$ instead of $\sup_{x \in S} |Q_{n,h}(x)| = O\{r(n,h)\}$. It follows from (3.7) that

$$
E\hat{\Psi}_h(x, \theta_\infty) = O \left( \frac{1}{n^{(1+\alpha)/2} h^2} \right).
$$

(A.1)

This implies

$$
\theta_{h}(x) = \theta_\infty - \left[ \frac{\partial}{\partial \theta} E\hat{\Psi}_h(x, \theta) \right]^{-1}_{\theta = \theta_\infty} E\hat{\Psi}_h(x, \theta_\infty) + O \left( \frac{1}{n^{(1+\alpha)/2} h^2} \right).
$$

(A.2)

Using the conditions (A2) and (A3), the identity (2.6), and the fact $\int \tilde{f}(x, \theta) \, dx = 0$, we may verify

$$
\left[ \frac{\partial}{\partial \theta} E\hat{\Psi}_h(x, \theta) \right]_{\theta = \theta_\infty} = -I_h(x, \theta_\infty) + O \left( \frac{1}{n^{(1+\alpha)/2} h^2} \right)
$$

(A.3)

$$
= -I_\infty(\theta_\infty) + O \left( \frac{1}{h^2} \right).
$$

Plug the second approximation at (A.3) into (A.2) and use (A.1) to get

$$
\theta_{h}(x) = \theta_\infty + I_\infty(\theta_\infty)^{-1} E\hat{\Psi}_h(x, \theta_\infty) + O(\rho).
$$

(A.4)

The lemma follows immediately from (A.4).
A.3. Proof of Lemma 3. In this proof, all \( \mathcal{O}_p \) expressions are also uniform for \( x \) in any compact subset \( S \) of \( \mathcal{X} \). First, we observe

\[
\hat{\theta}_h(x) = \theta_h(x) - \left[ \frac{\partial}{\partial \theta} \mathbb{E} \hat{\Psi}_h(x, \theta) \right]_{\theta = \theta_h(x)}^{-1} \hat{\Psi}_h(x, \theta_h(x)) + O_p \left( \frac{\log n}{n} \right). \tag{A.5}
\]

The proof of (A.5) is similar to that of (4.1) in Park et al. (2002). The only difference is that we let \( h \) tend to infinity instead of zero and thus only have \( O_p(n^{-1/2} \log n) \) instead of \( O_p(n^{-1} \log n) \) for the remainder. Now, we can replace \( \hat{\Psi}_h(x, \theta_h(x)) \) by \( U_h(x, \theta_h(x)) \) in (A.5) since \( E \hat{\Psi}_h(x, \theta_h(x)) = 0 \) by definition of \( \theta_h(x) \). Also, we can replace \( -\left[ \frac{\partial}{\partial \theta} E \hat{\Psi}_h(x, \theta) \right]_{\theta = \theta_h(x)} \) by \( I_h(x, \theta_\infty) \) with an error \( O_p(\{n^{-1-(\alpha/2)}h^{-2}(\log n)^{1/2}\}) \). This is due to the facts that \( E \hat{\Psi}_h(x, \theta_h(x)) = O_p(n^{-1/2} \log n) \) and that \( \left[ \frac{\partial}{\partial \theta} E \hat{\Psi}_h(x, \theta) \right]_{\theta = \theta_h(x)} - \left[ \frac{\partial}{\partial \theta} E \hat{\Psi}_h(x, \theta) \right]_{\theta = \theta_\infty} \) has magnitude of order \( O(n^{-(1+\alpha)/2}h^{-2}) \) by (A.1) and (A.2). Also, \( \left[ \frac{\partial}{\partial \theta} E \hat{\Psi}_h(x, \theta) \right]_{\theta = \theta_\infty} + I_h(x, \theta_\infty) = O(n^{-(1+\alpha)/2}h^{-2}) \) by the first approximation at (A.3). This yields

\[
\hat{\theta}_h(x) = \theta_h(x) + I_h(x, \theta_\infty)^{-1} U_h(x, \theta_h(x)) + O_p(\delta).
\]

The lemma then follows immediately from the facts \( U_h(x, \theta_\infty) = O_p(n^{-1/2} \log n) \) and \( \theta_h(x) - \theta_\infty = O(n^{-(1+\alpha)/2}h^{-2}) \). ■
Acknowledgments. We are grateful for the helpful and constructive comments of an associate editor and two reviewers.
References


