

# Powerful Choices: Tuning Parameter Selection Based on Power

Kjell A. Doksum\* and Chad M. Schafer†

*University of Wisconsin, Madison and Carnegie Mellon University*

October 13, 2005

## Abstract

We consider procedures which select the bandwidth in local linear regression by maximizing the limiting power for local Pitman alternatives to the hypothesis that  $\mu(x) \equiv E(Y | X = x)$  is constant. The focus is on achieving high power near a covariate value  $x_0$  and we consider tests based on data with  $X$  restricted to an interval containing  $x_0$  with bandwidth  $h$ . The power optimal bandwidth is shown to tend to zero as sample size goes to infinity if and only if the sequence of Pitman alternatives is such that the length of the interval centered at  $x_0$  on which  $\mu(x) = \mu_n(x)$  is nonconstant converges to zero as  $n \rightarrow \infty$ . We show that tests which are based on local linear fits over asymmetric intervals of the form  $[x_0 - (1 - \lambda)h, x_0 + (1 + \lambda)h]$ , where  $-1 \leq \lambda \leq 1$ , rather than the symmetric intervals  $[x_0 - h, x_0 + h]$  will give better asymptotic power. A simple procedure for selecting  $h$  and  $\lambda$  consists of using order statistics intervals containing  $x_0$ . Examples illustrate that the effect of these choices are not trivial: Power optimal bandwidth can give much higher power than bandwidth chosen to minimize mean squared error. Because we focus on power, rather than plotting estimates of  $\mu(x)$  we plot a correlation curve  $\widehat{\rho}(x)$  which indicates the strength of the dependence between  $Y$  and  $X$  near each  $X = x$ . Extensions to more general hypotheses are discussed.

## 1 Introduction. Local Testing and Asymptotic Power.

We consider  $(X_1, Y_1), (X_2, Y_2), \dots, (X_n, Y_n)$  i.i.d. as  $(X, Y) \sim P$  and write  $Y = \mu(X) + \epsilon$  where  $\mu(X) \equiv E(Y | X)$  and  $\beta(x) \equiv d\mu(x)/dx$  are assumed to exist, and  $\epsilon \equiv Y - \mu(X)$ . We are interested in a particular covariate value  $x_0$  and ask if there is a relationship between the response  $Y$  and the covariate  $X$  for  $X$  in some neighborhood of  $x_0$ . More formally, we test the hypothesis  $H$  that

---

\*Supported in part by NSF grants DMS-9971309 and DMS-0505651.

†Supported in part by NSF grant DMS-9971301.

$\beta(x) = 0$  for all  $x$  against the alternative  $K$  that  $\beta(x) \neq 0$  for  $x$  in some neighborhood of  $x_0$ . Our focus is on achieving high power for a covariate value  $x_0$  for a unit (patient, component, DNA sequence, etc.) of interest. We want to know if a perturbation of  $x$  will affect the mean response for units with covariate values near  $x_0$ .

Our test statistic is the  $t$ -statistic  $t_h(x_0)$  based on the locally linear estimate  $\widehat{\beta}_h(x_0)$  of  $\beta(x_0)$  obtained as the slope of the least squares estimate of  $\beta(x_0)$  computed for data in the local data set

$$D_h(x_0) \equiv \{(x_i, y_i) : x_i \in N_h(x_0)\}, \quad \text{with } N_h(x_0) \equiv [x_0 - h, x_0 + h].$$

The bandwidth  $h$  serves as a lense with different lenses providing insights into the relationship between  $Y$  and  $X$  (Chaudhuri and Marron (1999, 2000) did “estimation” lenses). The  $h$  that maximizes the power over  $N_h(x_0)$  provides a customized lense for a unit with covariate value  $x_0$ . We discuss global alternatives where  $\sup_x |\beta(x)| > 0$  in Remark 2.3. The many available tests for this alternative will have increased power if they are based on variable bandwidths.

We will next show that the power of the test is very sensitive to the choice of  $h$  and that this choice at first appears to be inversely related to the usual choice of  $h$  based on mean squared error (MSE). In fact, for any alternative with  $\beta(x)$  nonzero on an interval centered at  $x_0$  that does not shrink to a point as  $n$  goes to infinity, the asymptotic power is maximized by selecting an  $h = h_n$  that does not converge to zero as  $n \rightarrow \infty$ .

Because  $\widehat{\beta}_h(x_0)$  is asymptotically normal, the asymptotic power of the test based on it for local Pitman alternatives will be determined by its efficacy (EFF) for one-sided alternatives  $\beta(x) > 0$  and by its absolute efficacy for two-sided alternatives  $\beta(x) \neq 0$  (Kendall and Stuart (1961), Section 25.5, Lehmann (1999)):

$$\text{EFF}_h \equiv \text{EFF}(\widehat{\beta}_h(x_0)) = \text{E}_K[\widehat{\beta}_h(x_0)] / \left( \text{Var}_H[\widehat{\beta}_h(x_0)] \right)^{1/2}.$$

Consider the alternative  $\mu(x) = \alpha + \gamma r(x)$  for a differentiable function  $r(\cdot)$ . We know (e.g. Fan and Gijbels (1996), pg. 62) that for  $h \rightarrow 0$ , conditionally on  $\mathbf{X} = (X_1, X_2, \dots, X_n)$ ,

$$\text{E}_K[\widehat{\beta}_h(x_0) | \mathbf{X}] = \gamma r'(x_0) + c_1 h^2 + o_P(h^2) \quad \text{and}$$

$$\text{Var}_H[\widehat{\beta}_h(x_0) | \mathbf{X}] = c_2 n^{-1} h^{-3} + o_P(n^{-1} h^{-3})$$

for appropriate constants  $c_1$  and  $c_2$ . These expressions yield an approximate efficacy of order  $n^{1/2} h^{3/2}$  which is maximized by  $h$  not tending to zero (where the expressions are not valid). This

shows that  $h$  tending to zero does not yield optimal power and that maximizing power is typically very different from minimizing some variant of mean squared error: for example,

$$\text{MSE}(h) \equiv \mathbb{E} \left[ (\hat{\mu}_h(X) - \mu(X))^2 \mathbf{1}\{X \in (x_0 - h, x_0 + h)\} \right],$$

where  $\hat{\mu}_h(\cdot)$  is the local linear estimate of  $\mu(\cdot)$ . This subject has been discussed for global alternatives by many authors, sometimes with the opposite conclusion. For a discussion, see e.g. Hart (1997), Zhang (2003b), Zhang (2003a), and Stute and Zhu (2005). In the next section we look more closely at the optimal  $h$  and find that for certain least favorable distributions, the optimal  $h$  does indeed tend to zero.

**Remark 1.1:** Why the  $t$ -test?

(a) The  $t$ -test is asymptotically most powerful in the class of tests using data from  $D_h(x_0)$  for a class of perturbed normal linear models defined for  $x_i \in N_h(x_0)$  by  $Y_i = \gamma(x_i - x_0) + \gamma^2 r(x_i) + \epsilon$ , where  $\sum (x_i - x_0) = 0$ . Here,  $\epsilon$  has the distribution  $\Phi(t/\sigma_0)$  where  $\gamma = \gamma_n = O(n^{-1/2})$  is a Pitman alternative,  $h$  is fixed, and  $r(\cdot)$  is a general function satisfying regularity conditions. To establish the asymptotic power optimality, suppose all parameters are known except  $\gamma$  and consider the Rao score test for testing  $H: \gamma = 0$  versus  $K: \gamma > 0$ .

It is easy to see that this statistic is equivalent to  $\sum (x_i - x_0) Y_i$  whose asymptotic power is the same as that of the  $t$ -test. It is known (e.g. Bickel and Doksum (2001), Section 5.4.4) that the score test is asymptotically most powerful level  $\alpha$ . In the presence of nuisance parameters, a modified Rao score test retains many of its power properties, see Neyman (1959) and Bickel et al. (2006). The idea of asymptotic optimality for perturbed models is from Bell and Doksum (1966). The asymptotic power optimality of the  $t$ -test is in agreement with the result that when bias does not matter, the uniform kernel is optimal (e.g. Gasser et al. (1985) and Fan and Gijbels (1996)). But see Blyth (1993), who shows that for certain models, a kernel asymptotically equivalent to the derivative of the Epanechnikov kernel maximizes the power asymptotically.

While for alternatives of this type the use of a test based on the  $t$ -test is asymptotically most powerful, there undoubtedly exist other situations in which a test based on an appropriate non-parametric estimate of  $\mu(\cdot)$  gives better power. We will explore this issue in future work.

(b) The  $t$ -test is the natural model selector for choosing between the models  $\mu(x) = \beta_0 + \epsilon$  and  $\mu(x) = \beta_0 + \beta_1 x + \epsilon'$  in the sense that to minimize mean squared prediction error, we select the former model if the absolute value of the usual  $t$ -statistic is less than  $\sqrt{2}$  (see e.g. Linhart and

Zucchini (1986)). This is easily seen to be equivalent to minimizing the mean squared estimation error. It also is the rule of choice if we apply the methods of Claeskens and Hjorth (2003) who choose the model that minimizes the asymptotic mean squared estimation error for Pitman sequences of alternatives.

**Remark 1.2:** For a given test statistic  $T$ , the numerator of the efficacy is usually (e.g. Lehmann (1999), pg. 172, and Kendall and Stuart (1961)) defined in terms of the derivative of  $E_{\theta}(T)$  evaluated at the hypothesized parameter value  $\theta_0$ . We consider the “pre-derivative” because we want to investigate how  $h$  influences efficacy.

**Remark 1.3:** Hajek (1962), Behnen and Hušková (1984), Behnen and Neuhaus (1989), Sen (1996), among others, proposed and analyzed procedures for selecting the asymptotically optimal linear rank test by maximizing the estimated efficacy.

**Remark 1.4:** The approach of selecting bandwidth to maximize efficacy can also be used to test a parametric model of the form  $E(Y|X = x) = g(x; \beta)$  against a nonparametric model. Simply begin by subtracting off from  $Y$  the appropriate model fit under the (null hypothesis) parametric model. See Remark 2.9.

## 2 Maximizing Asymptotic Power

We start with fixed  $h > 0$ . Then the limiting power for the two-sided alternative is a one-to-one function of the absolute value of

$$\tau_h(x_0) \equiv \lim_{n \rightarrow \infty} n^{-1/2} \text{EFF}(\widehat{\beta}_h(x_0)).$$

Thus the maximizer of the asymptotic power is  $h_0 \equiv \arg \max_h |\tau_h(x_0)|$ . Our estimate of  $h_0$  is the maximizer of the absolute  $t$ -statistic,  $\widehat{h} \equiv \arg \max_h |t_h(x_0)|$  or its equivalents given in Remark 2.1. To investigate the limit of the  $t$ -statistic,  $t_h(x_0)$ , for the data  $D_h(x_0)$  we write it as signal/noise where

$$\begin{aligned} \text{signal} &\equiv \widehat{\beta}_h(x_0) = \widehat{\text{Cov}}_h(X, Y) / \widehat{\text{Var}}_h(X) \quad \text{and} \\ \text{noise}^2 &\equiv \widehat{\text{Var}}(\widehat{\beta}_h(x_0)) = \widehat{\text{E}}_h(\epsilon_h^2) / n_h \widehat{\text{Var}}_h(X). \end{aligned}$$

Here  $\widehat{\text{Var}}_h(X)$  and  $\widehat{\text{Cov}}_h(X, Y)$  denote the sample variance and covariance for the data  $D_h(x_0)$ ,

$$n_h \equiv \sum_{i=1}^n \mathbf{1}\{x_i \in N_h(x_0)\} \quad \text{and} \quad \widehat{\text{E}}_h(\epsilon_h^2) = \text{RSS}_h / (n_h - 2),$$

where  $\text{RSS}_h$  is the residual sum of squares  $\sum(y_i - \widehat{\mu}_{L,h}(x_i))^2$  for the linear fit  $\widehat{\mu}_{L,h}(x_i)$  to  $y_i$  based on  $D_h(x_0)$ .

For fixed  $h$ , as  $n \rightarrow \infty$ ,

$$\begin{aligned}\widehat{\beta}_h(x_0) &\xrightarrow{P} \beta_h(x_0) \equiv \text{Cov}_h(X, Y) / \text{Var}_h(X), \\ (n_h/n) &\xrightarrow{P} P_h(x_0) \equiv P(X \in N_h(x_0)), \quad \text{and} \\ \widehat{\text{E}}_h(\epsilon_h^2) &\xrightarrow{P} \text{E}_h(\epsilon_{L,h}^2) \equiv \text{E}_h[Y - \mu_{L,h}(X)]^2,\end{aligned}$$

where  $\text{E}_h$ ,  $\text{Var}_h$ , and  $\text{Cov}_h$  denote expected value, variance, and covariance conditional on  $X \in N_h(x_0)$ , and where  $\mu_{L,h}(x) = \alpha_h + \beta_h X$  with  $\alpha_h$  and  $\beta_h$  the minimizers of  $\text{E}_h[Y - (\alpha + \beta X)]^2$ . It follows that as  $n \rightarrow \infty$ ,

$$\begin{aligned}n^{-1/2} t_h(x_0) &\xrightarrow{P} \tau_h^*(x_0) \equiv \beta_h(x_0) \{P_h(x_0) \text{Var}_h(X)\}^{1/2} / \{\text{E}_h(\epsilon_{L,h}^2)\}^{1/2} \\ &= \left[ \text{SD}_H(Y) / (\text{E}_h(\epsilon_{L,h}^2))^{1/2} \right] \tau_h(x_0).\end{aligned}\tag{1}$$

For sequences of Pitman alternatives where  $\mu(x) = \mu_n(x)$  converges to a constant for all  $x$  as  $n \rightarrow \infty$ , we assume that as  $n \rightarrow \infty$

$$\text{E}_h(\epsilon_{L,h}^2) \rightarrow \sigma^2 \equiv \text{Var}_H(Y).$$

In this case,  $\tau_h^*(x_0) = \tau(x_0)$ . It is clear that  $\tau_h^*(x_0) \rightarrow 0$  as  $h \rightarrow 0$  because both  $P_h(x_0)$  and  $\text{Var}_h(X)$  tend to zero while  $\text{E}_h(\epsilon_{L,h}^2)$  does not tend to zero except in the trivial case  $Y = \mu_{L,h}(x)$  for  $x \in N_h(x_0)$ . On the other hand, if the density of  $X$  has support  $[a, b]$  and  $x_0 \in [a, b]$ , then for  $h' \equiv \max\{x_0 - a, b - x_0\}$ ,

$$\lim_{h \rightarrow h'} \tau_h^*(x_0) = \beta_L \left\{ \text{Var}(X) / \text{E}(\epsilon_L^2) \right\}^{1/2},$$

where  $\epsilon_L = Y - (\alpha_L + \beta_L X)$  with  $\alpha_L$  and  $\beta_L$  the minimizers of  $\text{E}[Y - (\alpha + \beta X)]^2$ , that is,  $\beta_L = \text{Cov}(X, Y) / \text{Var}(X)$ ,  $\alpha_L = \text{E}(Y) - \beta_L \text{E}(X)$ . Thus, when  $X$  has finite support  $[a, b]$  and  $x_0 \in [a, b]$ , the maximum of  $\tau_h^*(x_0)$  over  $h \in [0, h']$  exists and it is greater than zero. A similar argument shows that  $h = 0$  does not maximize  $\tau_h^*(x_0)$  when  $X$  has infinite support.

**Remark 2.1:** Instead of the  $t$ -statistic  $t_h(x_0)$ , we could use the efficacy estimate

$$\widehat{\text{EFF}}\left(\widehat{\beta}_h(x_0)\right) = r_h(x_0),$$

where  $r_h(x_0)$  is the sample correlation for data from  $D_h(x_0)$ . The formula

$$t_h^2(x_0) = (n - 2) r_h^2(x_0) / (1 - r_h^2(x_0)),$$

where the right hand side is increasing in  $|r_h(x_0)|$ , establishes the finite sample equivalence of  $|t_h(x_0)|$  and  $|r_h(x_0)|$ . They are also similar to  $|\rho(x_0)|$  where

$$\widehat{\rho}_h(x_0) = \frac{\widehat{\beta}_h(x_0) \widehat{\text{SD}}(X)}{\sqrt{\widehat{\beta}_h^2(x_0) \widehat{\text{Var}}(X) + \widehat{\text{Var}}_h(Y)}}$$

is the estimate of the correlation curve  $\rho(\cdot)$  discussed by Bjerve and Doksum (1993), Doksum et al. (1994), and Doksum and Froda (2000). Here  $\widehat{\rho}_h(x)$  is a standardized version of  $r_h(x_0)$  which is calibrated to converge to the correlation coefficient  $\rho$  in bivariate normal models.

## 2.1 Optimal Bandwidth for Vertical Pitman Sequences

If we consider Pitman alternatives of the form  $Y = \alpha + \gamma r(X) + \epsilon$  where  $\gamma = \gamma_n = c/\sqrt{n}$ ,  $c \neq 0$ , and  $|r'(x)| > 0$  in a fixed interval containing  $x_0$ , then

$$n^{-1/2} \text{EFF} \left[ \widehat{\beta}_h(x_0) \right] \rightarrow \tau_h(x_0),$$

where  $\tau_h(x_0)$  is as in Equation (1), and the optimal  $h$  will be bounded away from zero as before. Thus, the power optimal bandwidth does not tend to zero for sequences of alternatives where, as  $n \rightarrow \infty$ ,  $\|\beta\|_\infty \rightarrow 0$  and  $\|\beta\|^{-\infty} \not\rightarrow 0$  with  $\|\beta\|_\infty \equiv \sup_x |\beta(x)|$  and

$$\|\beta\|^{-\infty} \equiv x^+(\beta) - x^-(\beta) \equiv \sup_x \{x: |\beta(x)| > 0\} - \inf_x \{x: |\beta(x)| > 0\}.$$

We refer to  $\|\beta\|_\infty$  as the ‘‘vertical distance’’ between  $H$  and the alternative and  $\|\beta\|^{-\infty}$  as the ‘‘horizontal discrepancy.’’

Next we consider ‘‘horizontal’’ sequences of alternatives where  $\|\beta\|^{-\infty} \rightarrow 0$  and  $\|\beta\|_\infty$  may or may not tend to zero as  $n \rightarrow \infty$  and find that now the power optimal bandwidth tends to zero.

## 2.2 Optimal Bandwidth for Horizontal Pitman Sequences

We now consider Pitman alternatives of the form

$$K_n: Y = \alpha + \gamma W \left( \frac{X - x_0}{\theta} \right) + \epsilon, \tag{2}$$

where  $X$  and  $\epsilon$  are uncorrelated,  $\epsilon$  has mean zero and variance  $\sigma^2$ , and  $W(\cdot)$  has support  $[-1, 1]$ . We assume that  $X$  is continuous, in which case  $\mu(\cdot)$  is the constant  $\alpha$  with probability one when

$\theta = 0$ . Thus the hypothesis holds with probability one when  $\gamma\theta = 0$ . We consider models where  $\theta = \theta_n \rightarrow 0$  as  $n \rightarrow \infty$ , and  $\gamma$  may or may not depend on  $\theta$  and  $n$ . For these alternatives the neighborhood where  $|\beta(x)| > 0$  shrinks to achieve a Pitman balanced model where the power converges to a limit between the level  $\alpha$  and 1 as  $n \rightarrow \infty$ . Note, however, that the alternative does not depend on  $h$ . We are in a situation where “nature” picks the neighborhood size  $\theta$ , and the statistician picks the bandwidth  $h$ . This is in contrast to Blyth (1993) and Ait-Sahalia et al. (2001) who let the alternative depend on  $h$ .

Note that for  $h > 0$  fixed,  $\gamma$  bounded above, and  $\theta \rightarrow 0$ ,

$$\text{Cov}_h(X, Y) = \gamma \text{Cov}_h\left(X, W\left(\frac{X - x_0}{\theta}\right)\right) \rightarrow 0$$

because  $W((X - x_0)/\theta)$  tends to zero in probability as  $\theta \rightarrow 0$  and any random variable is uncorrelated with a constant. This heuristic can be verified by the change of variable  $s = (x - x_0)/\theta$ ,  $x = x_0 + \theta s$ . More precisely,

**Proposition 2.1:** Assume that the density  $f(\cdot)$  of  $X$  has a bounded derivative at  $x_0$ . If  $h > 0$  is fixed, then as  $\gamma\theta \rightarrow 0$  in the Model (2),

- (a)  $\text{Cov}_h(X, Y) = O(\gamma\theta)$ ;
- (b)  $n^{-1/2}\text{EFF}(\hat{\beta}_h(x_0)) \rightarrow 0$ ;
- (c)  $n^{-1/2}t_h(x_0) \xrightarrow{P} 0$ .

**Proof:** Since  $\epsilon$  is assumed uncorrelated with  $X$ ,  $\text{Cov}_h(X, Y) = \text{Cov}_h(X, \mu(X))$  and hence

$$\begin{aligned} \text{Cov}_h(X, Y) &= \gamma \text{Cov}_h\left(X, W\left(\frac{X - x_0}{\theta}\right)\right) \\ &= \gamma \int_{-h}^h (x - \text{E}_h(X)) W\left(\frac{x - x_0}{\theta}\right) \left[ f(x) / P_h(x_0) \right] dx \\ &= \frac{\gamma\theta}{P_h(x_0)} \int_{-1}^1 (x_0 + s\theta - \text{E}_h(X)) W(s) f(x_0 + s\theta) ds \\ &= \frac{\gamma\theta}{P_h(x_0)} \left[ (x_0 - \text{E}_h(X)) f(x_0) \int_{-1}^1 W(s) ds + O(\theta) \right]. \end{aligned} \tag{3}$$

Then Result (b) follows because

$$\text{E}_K\left(\hat{\beta}_h(x_0)\right) \rightarrow \text{Cov}_h(X, Y) / \text{Var}_h(X)$$

and the other factors in  $\text{EFF}_h$  are fixed as  $\theta \rightarrow 0$ . Similarly (c) follows from Equation (1).  $\square$

Proposition 2.1 shows that for model (2), fixed  $h$  leads to small  $|\text{EFF}_h|$ . Thus we turn to the  $h \rightarrow 0$  case. If  $h > \theta$  then observations  $(X, Y)$  with  $X$  outside  $[x_0 - \theta, x_0 + \theta]$  do not contribute to

the estimation of  $\beta_h(x_0)$ . Thus choosing a smaller  $h$  may be better even though a smaller  $h$  leads to a larger variance for  $\widehat{\beta}_h(x_0)$ . This heuristic is made precise in the next result which provides conditions where  $h = \theta$  is the optimal choice among  $h$  satisfying  $h \geq \theta$ .

Define

$$m_j(W) \equiv \int_{-1}^1 s^j W(s) ds \quad j = 0, 1, 2.$$

**Theorem 2.1:** Assume that the density  $f(\cdot)$  of  $X$  has a bounded, continuous second derivative at  $x_0$  and that  $f(x_0) > 0$ . Then, in model (2), as  $\theta \rightarrow 0$  and  $h \rightarrow 0$  with  $h \geq \theta$ , the following are true.

$$\begin{aligned} \text{(a) } \text{Cov}_h(X, Y) &= \frac{\gamma\theta}{2hf(x_0)} \left\{ -\frac{1}{3}m_0(W) f'(x_0) h^2 \right. \\ &\quad \left. + f(x_0) m_1(W) \theta + f'(x_0) m_2(W) \theta^2 + o(h^2) + o(\theta^2) \right\}. \end{aligned}$$

(b) If  $m_1(W) \neq 0$  and  $m_2(W) \neq 0$ ,

$$\begin{aligned} \tau_{h,\theta}(x_0) &\equiv \lim_{n \rightarrow \infty} n^{-1/2} \text{EFF}(\widehat{\beta}_h(x_0)) \\ &= \sigma^{-1} \gamma \theta h^{-3/2} [f(x_0)]^{-1/2} \left\{ -\frac{1}{3}m_0(W) f'(x_0) h^2 \right. \\ &\quad \left. + f(x_0) m_1(W) \theta + f'(x_0) m_2(W) \theta^2 + o(h^2) + o(\theta^2) \right\} \\ &= O\left(\sigma^{-1} \gamma \theta^2 h^{-3/2}\right) \end{aligned} \tag{4}$$

and  $|\tau_{h,\theta}(x_0)|$  is maximized subject to  $h \geq \theta$  by  $h = \theta$ .

(c) If  $m_1(W) = 0$ ,

$$\lim_{n \rightarrow \infty} n^{-1/2} \text{EFF}(\widehat{\beta}_h(x_0)) = O\left(\sigma^{-1} \gamma \theta^3 h^{-3/2}\right)$$

and  $|\tau_{h,\theta}(x_0)|$  is maximized subject to  $h \geq \theta$  by  $h = \theta$ .

**Proof:**

**Part (a):** This is Lemma 4.3, part (e).

**Part (b):**

$$\lim_{n \rightarrow \infty} n^{-1/2} \text{EFF}_h = \sigma^{-1} \text{Cov}_h(X, Y) \left[ (P_h(x_0)) / \text{Var}_h(X) \right]^{1/2}.$$

By Lemma 4.3, part (d), we get

$$P_h(x_0) / \text{Var}_h(X) = 6f(x_0) h^{-1} + O(1).$$

Equation (4) follows from this and part (a). Thus we can write

$$\lim_{n \rightarrow \infty} n^{-1/2} \text{EFF}_h \asymp \sigma^{-1} \left[ c_1 \gamma \theta h^{1/2} + c_2 \gamma \theta^2 h^{-3/2} + c_3 \gamma \theta^3 h^{-3/2} \right] \tag{5}$$

for appropriate constants  $c_1$ ,  $c_2$ , and  $c_3$  given in part (b). The square of the expression on the right of Equation (5) is dominated by  $[\sigma^{-1}\theta^2h^{-3/2}c_2]^2$  and maximized for  $h \geq \theta$  by  $h = \theta$ .

**Part (c):** Apply the argument in Part (b).  $\square$

It remains to investigate the case where  $h \leq \theta$ . We next show that the optimal  $h$  cannot be such that  $h/\theta \rightarrow 0$ .

**Proposition 2.2:** Assume the conditions of Theorem 2.1 and that  $W''(\cdot)$  exists and is bounded in a neighborhood of zero. Then, as  $h/\theta \rightarrow 0$ ,

$$\lim_{n \rightarrow \infty} n^{-1/2} \text{EFF}_h = \begin{cases} o(\sigma^{-1}\gamma\theta^2h^{-3/2}), & \text{if } W'(0) \neq 0 \\ o(\sigma^{-1}\gamma\theta^3h^{-3/2}), & \text{if } W'(0) = 0 \end{cases}.$$

**Proof:** Set  $a \equiv h/\theta$  and define

$$m_{j,a}(W) \equiv \int_{-a}^a s^j W(s) ds, \quad j = 0, 1, 2.$$

Equation (4) in Theorem 2.1, part (b), with  $m_j(W)$  replaced by  $m_{j,a}(W)$  is valid. A Taylor expansion gives

$$m_{1,a}(W) = [2W'(0) + aW''(0)] a^3 + o(a^3).$$

The result follows for the  $W'(0) \neq 0$  case. The other case is similar.  $\square$

The case where for the smallest optimal  $h$ ,  $h/\theta \rightarrow a_0$  as  $\theta \rightarrow 0$  with  $a_0 \in (0, 1)$  remains. It is possible that the optimal  $h$  satisfies this condition, for instance if  $W(\cdot)$  is steep on  $[-1/2, 1/2]$  and nearly constant otherwise. In this case Theorem 2.1 holds with  $m_j(W)$  replaced by  $m_{j,a}(W)$ .

The concept of efficacy is useful when  $\lim_{n \rightarrow \infty} \text{EFF}_h$  is finite because then the limiting power is between the level  $\alpha$  and one. From Equation (5) we see that when  $h = \theta$  and  $m_1(W) \neq 0$  this is the case when  $\gamma\theta^{1/2}\sigma^{-1} = cn^{-1/2}$  for some  $c > 0$ . We obtain the following corollary which follows by extending Theorem 2.1 to sequences of  $\gamma\theta^{1/2}/\sigma$ . For extensions to uniform convergence for curve estimates see Einmahl and Mason (2005).

**Corollary 2.1:** Assume model (2) and the conditions of Theorem 2.1 and Proposition 2.2.

(a) Suppose that  $m_{1,b}(W) \neq 0$  for some  $b$  in  $(0, 1]$ . Then there exists  $a_0 \in (0, 1]$  such that the smallest  $h$  that maximizes

$$\lim_{n \rightarrow \infty} n^{-1/2} \text{EFF}(\hat{\beta}_h(x_0))$$

satisfies  $h/\theta \rightarrow a_0$  for  $\theta \rightarrow 0$ . Moreover,  $m_{1,a_0}(W) \neq 0$ , and if  $\gamma\theta^{1/2}/\sigma = cn^{-1/2}$ , then

$$\sup_h \left[ \lim_{n \rightarrow \infty} \text{EFF}_h \right] = \sqrt{(3/2) f(x_0)} m_{1,a_0}(W) a_0^{-3/2} c + o(\gamma\theta^{1/2}/\sigma).$$

(b) The smallest optimal  $h$  in part (a) equals  $\theta$  if and only if  $a^{-3/2} m_{1,a}(W)$  is maximized subject to  $a \leq 1$  by  $a = 1$ .

(c) If  $m_{1,b}(W) = 0$  for all  $b \in (0, 1]$ , that is  $W(\cdot)$  is symmetric about zero, and if  $m_{2,b}(W) \neq 0$  for some  $b \in (0, 1]$ , then there exists  $a_0 \in (0, 1]$  such that the smallest  $h$  that maximizes

$$\lim_{n \rightarrow \infty} n^{-1/2} \text{EFF}(\widehat{\beta}_h(x_0))$$

satisfies  $h/\theta \rightarrow a_0$  for  $\theta \rightarrow 0$ . Moreover  $m_{2,a_0}(W) \neq 0$ , and if  $(\gamma\theta^{3/2}/\sigma) = cn^{-1/2}$ , then

$$\begin{aligned} \sup_h \left[ \lim_{n \rightarrow \infty} \text{EFF}_h \right] &= \sqrt{3/2} [f(x_0)]^{-1/2} f'(x_0) \{m_{2,a_0}(W) \\ &- m_{0,a_0}(W)/3\} a_0^{1/2} c + o(\gamma\theta^{3/2}/\sigma). \end{aligned}$$

**Remark 2.2:** Doksum (1966) considered minimax linear rank tests for models where  $Y \stackrel{\mathcal{L}}{=} X + V(X)$  with  $V(X) \geq 0$  and  $\|F_Y - F_X\|_\infty \geq \theta$  and found that the least favorable distribution for testing  $H: V(\cdot) = 0$  has  $Y = X + V_0(X)$  with  $X$  uniform(0, 1) and  $V_0(x) = [a + \theta - x]_+ \mathbf{1}\{x \geq \xi\}$ , with  $\xi \in [0, 1 - \theta]$ . He considered horizontal Pitman alternatives with  $\|V_0\|^{-\infty} = \theta \rightarrow 0$  as  $n \rightarrow \infty$ . Fan (1992), Fan (1993), and Fan and Gijbels (1996) considered minimax kernel estimates for models where  $Y = \mu(X) + \epsilon$  and found that the least favorable distribution for estimation of  $\mu(x_0)$  using asymptotic mean squared error has  $Y = \mu_0(X) + \epsilon$  with

$$\mu_0(x) = \frac{1}{2} b_n^2 \left[ 1 - c \left( \frac{x - x_0}{b_n} \right)^2 \right]_+$$

where  $b_n = c_0 n^{-1/5}$ , for some positive constants  $c$  and  $c_0$ . Again,  $\|\mu_0\|^{-\infty} \rightarrow 0$  as  $n \rightarrow \infty$ . Similar results were obtained for the white noise model by Donoho and Liu (1991a,b). Lepski and Spokoiny (1999), building on Ingster (1982), considered minimax testing using kernel estimates of  $\mu(x)$  and a model with normal errors  $\epsilon$  and  $\text{Var}(\epsilon) \rightarrow 0$  as  $n \rightarrow \infty$ . Their least favorable distributions (page 345) are random linear combinations of functions of the form

$$\mu_j(x) = \left( h^{1/2} \int W^2(t) dt \right)^{-1} W\left(\frac{x - t_j}{h}\right),$$

which seems to indicate a model that depends on  $h$ . Here each  $\mu_j(x)$  has  $\|\mu_j\|^{-\infty} \rightarrow 0$  as  $n \rightarrow \infty$ .

**Remark 2.3: Global Alternatives.** The idea of selecting  $h$  by maximizing the estimated efficacy of the  $t$ -statistic extends to global alternatives where the alternative is  $\|\mu(\cdot) - \mu_Y\|_\infty > 0$  or  $\|\beta\|_\infty > 0$ . Efficacies can be built from test statistics in the literature such as those based on ratios of residual sums of squares for the two models under consideration or on

$$\sum_{i=1}^n [\hat{\mu}_{2,h}(X_i) - \hat{\mu}_{1,h}(X_i)]^2 V(X_i),$$

where  $\hat{\mu}_{2,h}(\cdot)$  and  $\hat{\mu}_{1,h}(\cdot)$  and estimates of  $\mu(\cdot)$  for two models and  $V(\cdot)$  is a weight function (e.g. Azzalini et al. (1989), Raz (1990), Doksum and Samarov (1995), Hart (1997), Ait-Sahalia et al. (2001), Fan et al. (2001), Zhang (2003a)).

Note that varying bandwidths can also be used for these types of situations. In the case of

$$\sum_{i=1}^n [\hat{\mu}_{2,h}(X_i) - \mu_{1,h}(X_i)]^2 V(X_i),$$

we could think of

$$T_h(x_0) = [\hat{\mu}_{2,h}(x_0) - \hat{\mu}_{1,h}(x_0)] V^{1/2}(x_0)$$

as a local test statistic and select  $h$  to maximize its squared efficacy. For instance, for testing  $\beta(x) = 0$  versus  $\|\beta\|_\infty > 0$  using  $\hat{\mu}_{1,h}(X_i) = \bar{Y}$  and  $\hat{\mu}_{2,h}(X)$  a locally linear kernel estimate, the efficacy of  $T_h(x_0)$  with  $V(\cdot) = 1$  will be approximately the same as the efficacy we obtained from  $\hat{\beta}_h(x_0)$ . Let  $h(x_0)$  denote the maximizer of the squared efficacy of  $T_h(x_0)$  and set  $h_i = h(X_i)$ . Then the final test statistic would be

$$\sum_{i=1}^n [\hat{\mu}_{2,h_i}(X_i) - \bar{Y}]^2 V(X_i).$$

Fan et al. (2001) and Zhang (2003a) proposed a ‘‘multiscale’’ approach to the varying coefficient model where they select  $h$  by maximizing a standardized version of the generalized likelihood ratio (GLR) statistic which is defined as the logarithm of the ratio of the residual sums of squares for the two models under consideration. The standardization consists of subtracting the null hypothesis asymptotic mean and dividing by the null hypothesis asymptotic standard deviation. That is, for situations where the degrees of freedom of the asymptotic chi-square distribution of the GLR statistic tends to infinity they are maximizing the efficacy of the GLR statistic.

Other extensions would be to select  $h$  to maximize

$$T_{h,\lambda}(\mathbf{x}_0) = \sum_{j=1}^g t_{h,\lambda}^2(x_{0j}) / g \quad \text{or}$$

$$T_{h,\lambda}^*(\mathbf{x}_0) = \sum_{i=1}^g \widehat{\beta}_{h,\lambda}^2(x_{0i}) / \text{SE} \left( \sum_{j=1}^g \widehat{\beta}_{h,\lambda}^2(x_{0j}) \right)$$

where  $\mathbf{x}_0 \equiv (x_{01}, x_{02}, \dots, x_{0g})^T$  is a vector of grid points. Finally, we could select  $h$  by, for some weight function  $v(\cdot)$ , maximizing the integrated efficacy  $\text{IEFF} \equiv \int \text{EFF}(x) v(x) dx$ .

**Remark 2.4:** Hall and Heckman (2000) used the maximum of local  $t$ -statistics to test the global hypothesis that  $\mu(\cdot)$  is monotone. Their estimated local regression slope is the least squares estimate based on  $k$  nearest neighbors and the maximum is over all intervals with  $k$  at least 2 and at most  $m$ . They established unbiasedness and consistency of their test rule under certain conditions.

**Remark 2.5:** Hall and Hart (1990) found that for a nonparametric two sample problem with null hypothesis of the form  $H_0 : \mu_1(x) = \mu_2(x)$  for all  $x$ , a test statistic which is a scaled version of  $n^{-1} \sum [\widehat{\mu}_{2h}(X_i) - \widehat{\mu}_{1h}(X_i)]^2$  has good power properties when the bandwidths in  $\widehat{\mu}_{jh}(x), j = 1, 2$  satisfy  $h/n \rightarrow p > 0$  as  $n \rightarrow \infty$  provided

$$s(\mu_1, \mu_2) \equiv \int_0^1 \left\{ \int_t^{t+p} [\mu_2(x) - \mu_1(x)] f(x) dx \right\}^2 dt > 0$$

when  $f(x)$  has support  $(0, 1)$ . Note that the asymptotic behavior of  $n^{1/2} s^{1/2}(\mu_1, \mu_2)$  for sequences of Pitman alternatives depends completely on whether the length of the set on which  $[\mu_2(x) - \mu_1(x)]$  differs from zero tends to zero as  $n \rightarrow \infty$ , just as in our case, where the limiting behavior of the scaled efficacy depends on whether the length of  $\{x : |\beta(x)| > 0\}$  tends to zero.

**Remark 2.6:** Godtliebsen et al. (2004) consider two-dimensional  $\mathbf{X}$  and use the sum of two squared local directional  $t$ -statistics. We study optimal bandwidths for this case in a subsequent paper.

### 2.3 Asymmetric Windows

Section 2.2 shows that the limiting power depends crucially on  $m_1(W)$  and on  $m_{1,a}(W)$ . If these are zero, the limiting efficacy is of smaller order than if they are not. A simple method for increasing the limiting power is to use locally linear methods over asymmetric windows of the form

$$N_{h,\lambda}(x_0) \equiv [x_0 - (1 - \lambda)h, x_0 + (1 + \lambda)h], \quad h > 0, \quad -1 \leq \lambda \leq 1.$$

Now the test statistic is the slope  $\widehat{\beta}_{h,\lambda}(x_0)$  of the least squares line for the data  $D_{h,\lambda}(x_0) \equiv \{(x_i, y_i) : x_i \in N_{h,\lambda}(x_0)\}$ . The efficacy  $\text{EFF}_{h,\lambda}$  will be computed conditionally on  $X \in N_{h,\lambda}(x_0)$  and in model (2) with  $h/\theta = a \in (0, 2]$  the expansion of  $\text{Cov}_{h,\lambda}(X, Y)$  will be in terms of the

integrals

$$m_{j,a,\lambda}(x_0) \equiv \begin{cases} \int_{-(1-\lambda)a}^{(1+\lambda)a} s^j W(s) ds, & \text{if } (1+\lambda)a \leq 1, (1-\lambda)a \leq 1 & (6) \\ \int_{-(1-\lambda)a}^1 s^j W(s) ds, & \text{if } (1+\lambda)a > 1, (1-\lambda)a \leq 1 & (7) \\ \int_{-1}^{(1+\lambda)a} s^j W(s) ds, & \text{if } (1+\lambda)a \leq 1, (1-\lambda)a > 1 & (8) \end{cases}$$

If  $W(\cdot)$  is symmetric, the term of order  $(\gamma\theta^2/\sigma)h^{-3/2}$  in Equation (5) will not vanish and we will maintain high test efficiency. Even if  $W(\cdot)$  is not symmetric, absolute efficacy is always increased because

$$\sup_{h,\lambda} |\text{EFF}_{h,\lambda}| \geq \sup_h |\text{EFF}_{h,0}|.$$

Note that with asymmetric windows we need  $h \geq 2\theta$  to make sure the intervals  $N_{h,\lambda}(x_0)$  cover the support of  $W(\cdot)$ .

**Theorem 2.2:** Assume that the density  $f(\cdot)$  of  $X$  has a bounded, continuous second derivative at  $x_0$  and that  $f(x_0) > 0$ . Also assume that we can find  $\lambda \in [0, 1]$  such that  $m_{j,0,\lambda}(W) \neq 0$  for either  $j = 0$  or  $j = 1$ . Then, in model (2), as  $\theta \rightarrow 0$  and  $h \rightarrow 0$  with  $h \geq 2\theta$ , the following are true.

$$(a) \quad \text{Cov}_{h,\lambda}(X, Y) = \frac{\gamma\theta}{2h} \{m_{1,a,\lambda}(W)\theta - m_{0,a,\lambda}(W)\lambda h + o(h) + o(\theta)\}.$$

$$(b) \quad \tau_{h,\theta,\lambda}(x_0) \equiv \lim_{n \rightarrow \infty} n^{-1/2} \text{EFF}(\hat{\beta}_{h,\lambda}(x_0)) = \sqrt{3/2}\sigma^{-1}\gamma\theta h^{-3/2} [f(x_0)]^{1/2} \\ \times \{m_{1,a,\lambda}(W)\theta - m_{0,a,\lambda}(W)\lambda h + o(h) + o(\theta)\}$$

and  $|\tau_{h,\theta}(x_0)|$  is maximized subject to  $h \geq 2\theta$  by  $h = 2\theta$ .

**Proof: Part (a):** In Lemma 4.5, part (b), we show that

$$P_{h,\lambda}(x_0) = 2f(x_0)h + 2f'(x_0)\lambda h^2 + o(h^2) \quad \text{and}$$

$$E_{h,\lambda}(X) = x_0 + \lambda h + [f'(x_0)/f(x_0)]h^2/3 + o(h^2).$$

Other terms are from Equation (3) with ranges of integration in Equations (6), (7), and (8), rather than from -1 to 1; and the proof of Lemma 4.3, part (e).

**Part (b):**

$$\lim_{n \rightarrow \infty} n^{-1/2} \text{EFF}_h = \sigma^{-1} \text{Cov}_{h,\lambda}(X, Y) \left[ (P_{h,\lambda}(x_0)) / \text{Var}_{h,\lambda}(X) \right]^{1/2}.$$

By Lemma 4.5, part (d), we get

$$P_{h,\lambda}(x_0) / \text{Var}_{h,\lambda}(X) = 6f(x_0)h^{-1} + O(1).$$

Equation (4) follows from this and part (a).  $\square$

**Corollary 2.2:** Under the conditions of Theorem 2.2, for the model (2), and  $\theta \rightarrow 0$ , then, if  $\gamma\theta^{1/2}/\sigma = cn^{-1/2}$ ,

$$\sup_{h \geq 2\theta} \left[ \lim_{n \rightarrow \infty} \text{EFF}_h \right] = \sqrt{3/2} [f(x_0)]^{1/2} [m_{1,a,\lambda}(W) - 2m_{0,a,\lambda}(W)\lambda] c + o(\theta).$$

The optimal  $h$  cannot be such that  $h/\theta \rightarrow 0$ .

**Proposition 2.3:** Assume the conditions of Theorem 2.2 with  $\lambda \in (0, 1)$  and that  $W''(\cdot)$  exists and is bounded in a neighborhood of zero. Then, as  $h/\theta \rightarrow 0$ ,

$$\lim_{n \rightarrow \infty} n^{-1/2} \text{EFF}(\hat{\beta}_h(x_0)) = o\left(\sigma^{-1} \gamma \theta h^{-3/2} \{m_{1,a,\lambda}(W) \theta - m_{0,a,\lambda}(W) \lambda h\}\right).$$

**Proof:** The proof is similar to the proof of Proposition 2.2.  $\square$

**Corollary 2.3:** Assume model (2) and the conditions of Theorem 2.2 and Proposition 2.3. Then there exists  $a_0 \in (0, 2]$  such that the smallest  $h$  that maximizes

$$\lim_{n \rightarrow \infty} n^{-1/2} \text{EFF}_{h,\lambda}(\hat{\beta}_{h,\lambda}(x_0))$$

satisfies  $h/\theta \rightarrow a_0$  for  $\theta \rightarrow 0$ . Moreover, if  $\gamma\theta^{1/2}/\sigma = cn^{-1/2}$ , then

$$\begin{aligned} \sup_h \left[ \lim_{n \rightarrow \infty} \text{EFF}_{h,\lambda} \right] &= \sqrt{(3/2) f(x_0)} [m_{1,a_0,\lambda}(W) - m_{0,a_0,\lambda}(W) a_0 \lambda] \\ &\times a_0^{-3/2} c + o\left(\gamma\theta^{1/2}/\sigma\right). \end{aligned}$$

### Remark 2.7: Computations

We select  $h$  and  $\lambda$  to maximize the absolute  $t$ -statistic  $|t_{h,\lambda}(x_0)|$  computed for the data from  $D_{h,\lambda}(x_0)$ . This is not a challenge computationally since this is a search for the largest absolute  $t$ -statistic over **all** possible intervals that include  $x_0$ . We approximate this search by using all order statistic intervals of the form  $(x_{(i)}, x_{(j)})$  that contain  $x_0$  and at least  $k$  points from  $\{x_1, x_2, \dots, x_n\}$  where  $k \approx 0.05n$  is a good choice.

### Remark 2.8: Critical Values

Under the hypothesis that  $X$  and  $Y$  are independent, we get a distribution free critical value by using the permutation distribution of the maximum absolute  $t$ -statistic obtained by maximizing  $t$ -statistics over local neighborhoods. By permuting  $(Y_1, Y_2, \dots, Y_n)$ , leaving  $(X_1, X_2, \dots, X_n)$  fixed

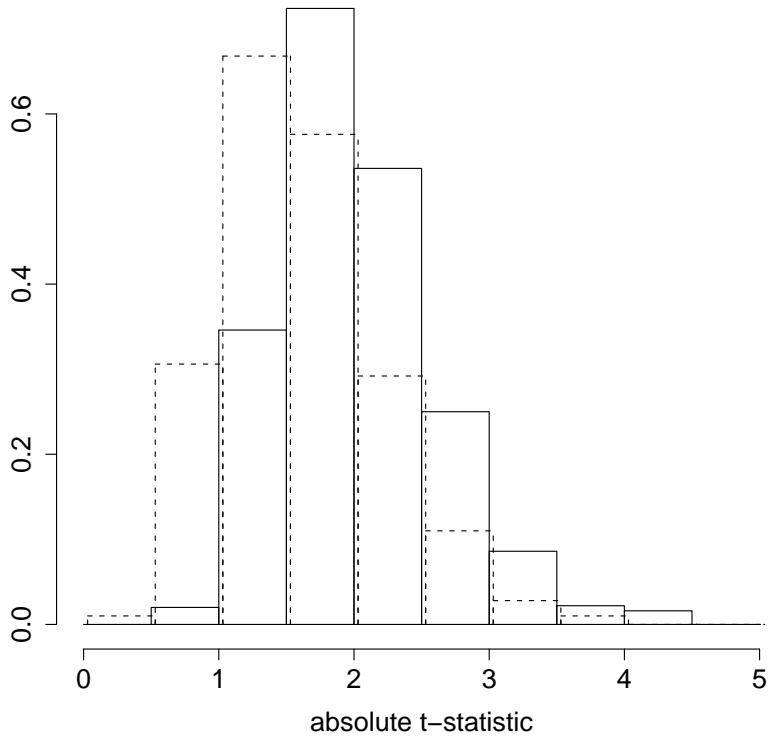


Figure 1: Approximate distribution of maximum absolute  $t$ -statistic under the null hypothesis, when using symmetric neighborhoods (dashed) and asymmetric neighborhoods (solid).

and computing the maximum  $t$ -statistic for these permuted data  $(X_i, Y_i), 1 \leq i \leq n$ , we get  $n!$  equally likely (under  $H$ ) values of the maximum absolute  $t$ -statistic. A subset of the  $n!$  permutations are chosen at random to reduce computational complexity. The dashed histogram in Figure 1 shows the approximated distribution using neighborhoods centered at  $x_0$ . We use 1000 random permutations. For each permutation of the data set, a set of ten bandwidths are used, ranging from the smallest which will ensure at least 20 data points in the neighborhood up to a bandwidth which selects all available data. The solid histogram in Figure 1 shows the approximated distribution using asymmetric neighborhoods which include  $x_0$ . The set of candidate endpoints of the neighborhoods are the quantiles  $\hat{x}_\alpha = \hat{F}^{-1}(\alpha)$  of the observed  $X$  variable, with  $\alpha$  uniformly spaced from zero up to one. These lead to 45 total neighborhoods, not all of which will include  $x_0$ . Since it searches over a larger class of neighborhoods, it is not surprising that the simulated distribution using asymmetric neighborhoods has a somewhat larger right tail than that for symmetric intervals.

**Remark 2.9: Testing a Parametric Hypothesis**

Suppose that we want to test a linear model against the alternative that near  $x_0$  a nonlinear model leads to a larger signal to noise ratio. Thus, near  $x_0$ , a local  $t$ -statistic may be significant when the global  $t$ -statistic is not, or visa versa. If we introduce  $Y'_i \equiv Y_i - (\hat{\alpha} + \hat{\beta}X_i)$  where  $\hat{\alpha} + \hat{\beta}x$  is the global least squares fit, and if we have in mind horizontal alternatives where the optimal  $h$  tends to zero, the previous results apply because  $(\hat{\alpha} - \alpha)$  and  $(\hat{\beta} - \beta)$  will converge to zero at a faster rate than  $(\hat{\beta}_h(x_0) - \beta_h(x_0))$  and thus testing a linear model for  $Y_i$  is asymptotically equivalent to testing that  $E(Y'|X = x)$  is constant. More generally, we may want to test a parametric model of the form  $E(Y|X = x) = g(x; \beta)$  against the alternative that we get higher power near  $x_0$  by using a nonparametric model. If we set  $Y' = Y_i - g(x; \hat{\beta})$  for suitable  $\sqrt{n}$  consistent  $\hat{\beta}$  and smooth  $g(\cdot)$ , the remark about the linear hypothesis still applies.

**3 Examples**

This section illustrates the behavior of the bandwidth selection procedure based on maximizing absolute efficacy, and compare it with methods based on MSE, using simulated and real (currency exchange) data sets. We utilize the model

$$\mu_\nu(x) = \nu x + \exp\left(-40(x - 0.5)^2\right), \quad 0 \leq x \leq 1,$$

and focus on the “flat bump” ( $\nu = 0$ ) and “sloped bump” ( $\nu = 1.25$ ) models shown in Figure 2. Assume  $X_1, X_2, \dots, X_n$  are uniform on  $(0, 1)$ , and that  $Y_i - \mu(X_i)$  is normal with mean zero and variance  $\sigma^2$ . Given a sample of data, MSE will be approximated using leave-one-out cross validation.

**3.1 Symmetric Windows**

The left plot in Figure 3 shows both theoretical  $MSE(h)$  and limiting efficacy ( $|\tau_h(x_0)|$ ) for a range of  $h$  in the sloped bump model, with  $x_0 = 0.6$ ,  $n = 1000$ , and  $\sigma^2 = 0.05$ . The mean squared error is minimized by choosing  $h = 0.07$ , while  $|\tau_h(x_0)|$  is maximized when  $h = 0.6$ . This discrepancy is to be expected: Bias of the local linear model estimator for  $\mu(x)$  is a critical component of the mean squared error, and the linear fit clearly becomes quickly inappropriate for  $x$  outside of the range  $[0.53, 0.67]$ . Alternatively, imagine one were to ask: “What interval centered at 0.6 would give the best chance of rejecting the hypothesis that  $\mu(x)$  is constant?” For sufficiently large  $\sigma^2$ , the

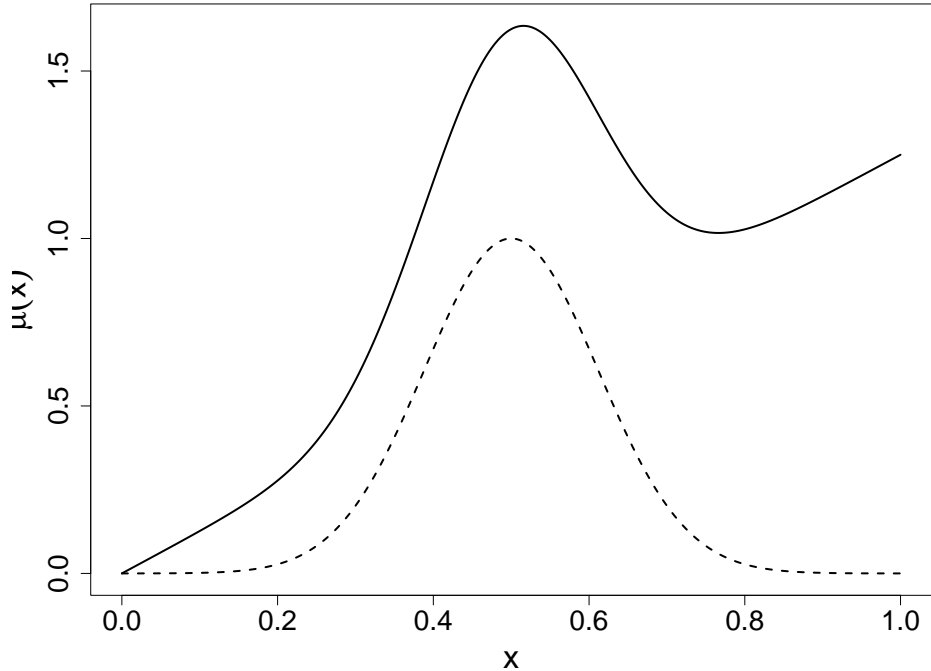


Figure 2: The “sloped bump” (solid) and “flat bump” (dashed) models.

procedure tells us to use the entire range of data for the most power to reject the hypothesis that  $\mu(\cdot)$  is constant. Here, bias is not at issue, we are instead searching for prevalent linear features in the bivariate relationship. There is a large drop in power if we use the MSE optimal  $h$  rather than the power optimal bandwidth.

There is, however, a local maximum in the left plot near  $h = 0.18$ . Switching to the flat bump model, we see that the chosen bandwidth is indeed  $h \approx 0.18$ ; see the right plot of Figure 3. The point is the following: When the overall linear trend is present and  $\sigma^2 = 0.05$ , the local downslope of the bump was not a sufficiently significant feature relative to the overall slope. Once  $\sigma^2 < 0.008$ , the power optimal bandwidth is chosen small. This illustrates the different behavior for “vertical” and “horizontal” alternatives.

The optimal bandwidth as chosen to maximize  $|\tau_h(x_0)|$  does not depend on  $n$ , but it does depend on  $\sigma^2$ . This is appropriate given  $\sigma^2$  is a feature of the bivariate distribution of  $X$  and  $Y$ , while  $n$  is not. Using the flat bump model, Figure 4 shows how the theoretically optimal bandwidth chosen by both minimizing MSE and maximizing  $|\tau_h(x_0)|$  varies with sample size, and Figure 5 shows the same as  $\sigma^2$  varies. Note that as  $\sigma^2$  increases, the power optimal bandwidth also grows. Figures

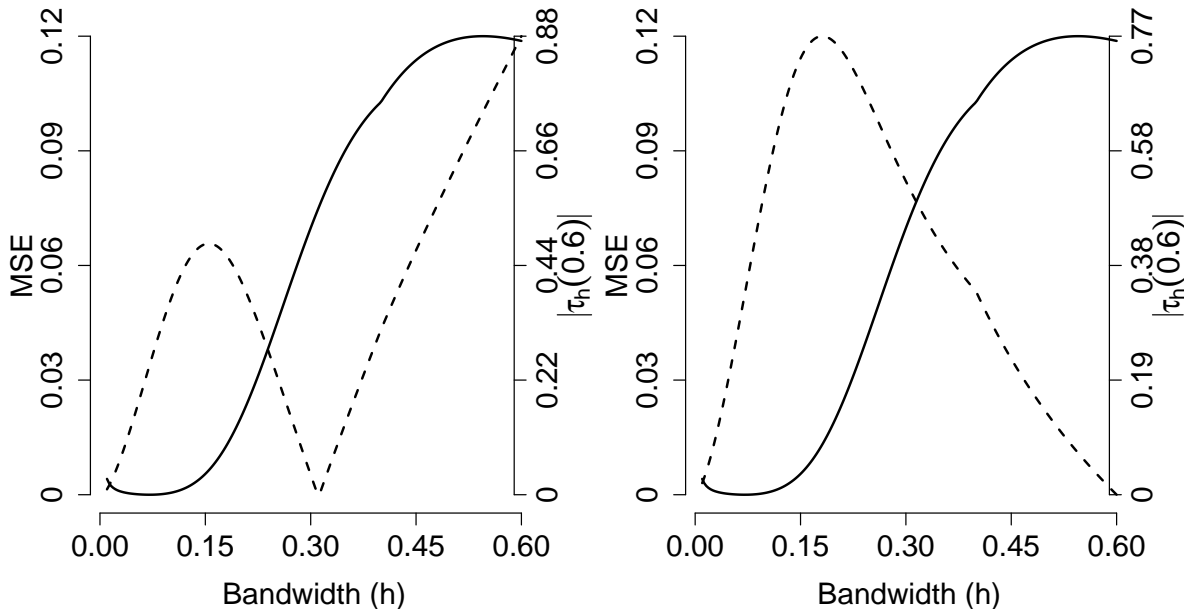


Figure 3: Plot of MSE (solid) and  $|\tau_h(x_0)|$  (dashed) for  $x_0 = 0.6$ ,  $n = 1000$ , and  $\sigma^2 = 0.05$  when using the sloped bump model (left plot) and the flat bump model (right plot).

4 and 5 also depict the bandwidth selection procedures on simulated data sets. A pair of dots connected by a vertical dotted line give the bandwidth selected by minimizing the leave-one-out cross validation (filled circle) and by maximizing the absolute  $t$ -statistic ( $|t_h(x_0)|$ ) (open circle), using the same simulated data set.

### 3.2 Asymmetric Windows

We now construct the optimal asymmetric windows, as first mentioned in Remark 2.7. Adjacent points along the  $x$  axis will often “share” the same optimal neighborhood. Figure 6 shows the result of maximizing  $|t_{h,\lambda}(x_0)|$  for each value of  $x_0$ . For this case, we use the sloped bump model with  $n = 1000$  and  $\sigma^2 = 0.05$ . For any chosen  $x_0$ , one can read up to find the neighborhood for that  $x_0$  by finding black horizontal line segment which intersects that vertical slice. Once the black segment is found, however, note that the corresponding neighborhood is in fact the **entire** length of the line segment, both the black and gray portions. For example, if  $x_0 = 0.75$ , the power optimal neighborhood extends from zero up to approximately 0.88. Figure 7 depicts the neighborhoods chosen in the same way, except now based on maximizing the absolute  $t$ -statistic when using a

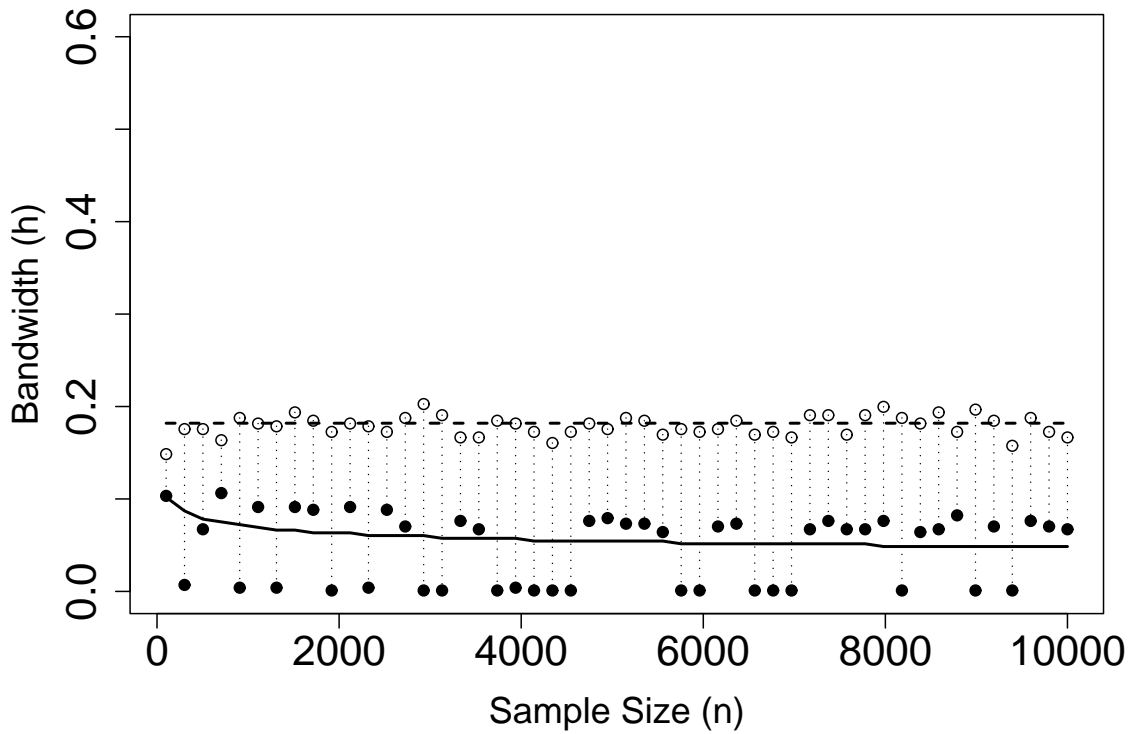


Figure 4: Plot of bandwidth found by minimizing theoretical MSE (solid) and maximizing theoretical  $|\tau_h(x_0)|$  (dashed) for  $x_0 = 0.6$  and  $\sigma^2 = 0.05$  when using the flat bump model. The dots represent the chosen bandwidth based on maximizing the absolute  $t$ -statistic (open) and minimizing the leave-one-out cross validation MSE (filled) using a randomly generated data set of the given size.

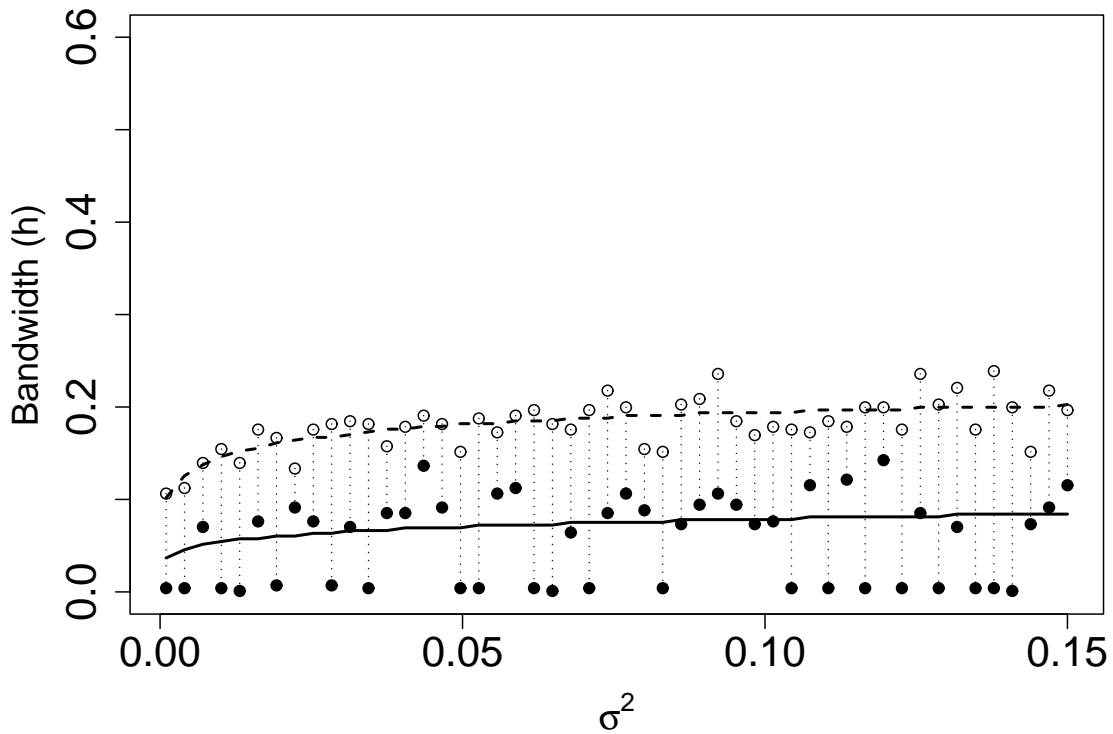


Figure 5: Plot of bandwidth found by minimizing theoretical MSE (solid) and maximizing theoretical  $|\tau_h(x_0)|$  (dashed) for  $x_0 = 0.6$  and  $n = 1000$  when using the flat bump model. The dots represent the chosen bandwidth based on maximizing the absolute  $t$ -statistic (open) and minimizing the leave-one-out cross validation MSE (filled) using a randomly generated data set of the given error variance ( $\sigma^2$ ).

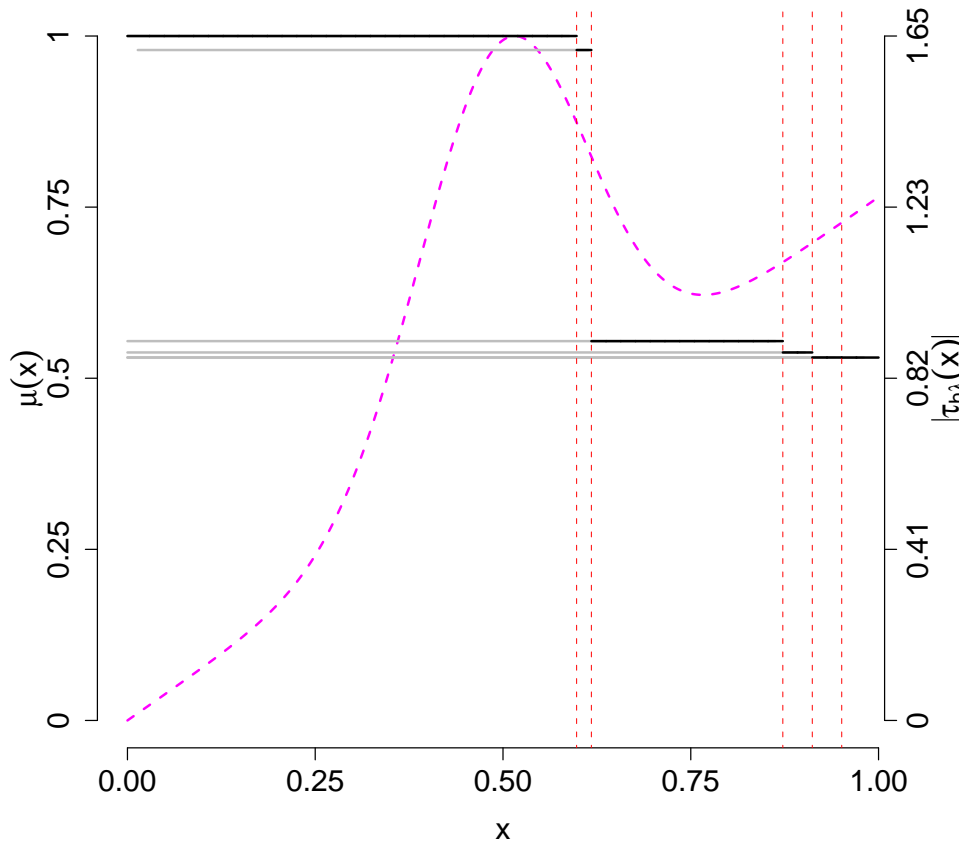


Figure 6: Plot of bandwidth chosen by maximizing  $|\tau_{h,\lambda}(x_0)|$  for all  $x_0$ , for the case  $n = 1000$ , and  $\sigma^2 = 0.05$  when using the sloped bump model.

simulated data set of size  $n = 1000$ .

### 3.3 Application to Currency Exchange Data

Figure 8 shows the results of this approach applied to daily Japanese Yen to Dollar currency exchange rate data for the period January 1, 1992 to April 16, 1995. The variable along the  $X$  axis is the standardized return, i.e. the logarithm the ratio of today's exchange rate to the yesterday's rate, standardized to have mean zero and variance one. The response variable is the logarithm of the volume of currency exchange for that day. Figure 9 displays, as the solid line, the values of  $\hat{\mu}_{h,\lambda}(x_0)$  for all values of  $x_0$ , when the bandwidth is chosen by maximizing  $|t_{h,\lambda}(x_0)|$ . The plot also shows, as a dashed line, an estimate of formed using local linear regression with tri-cube

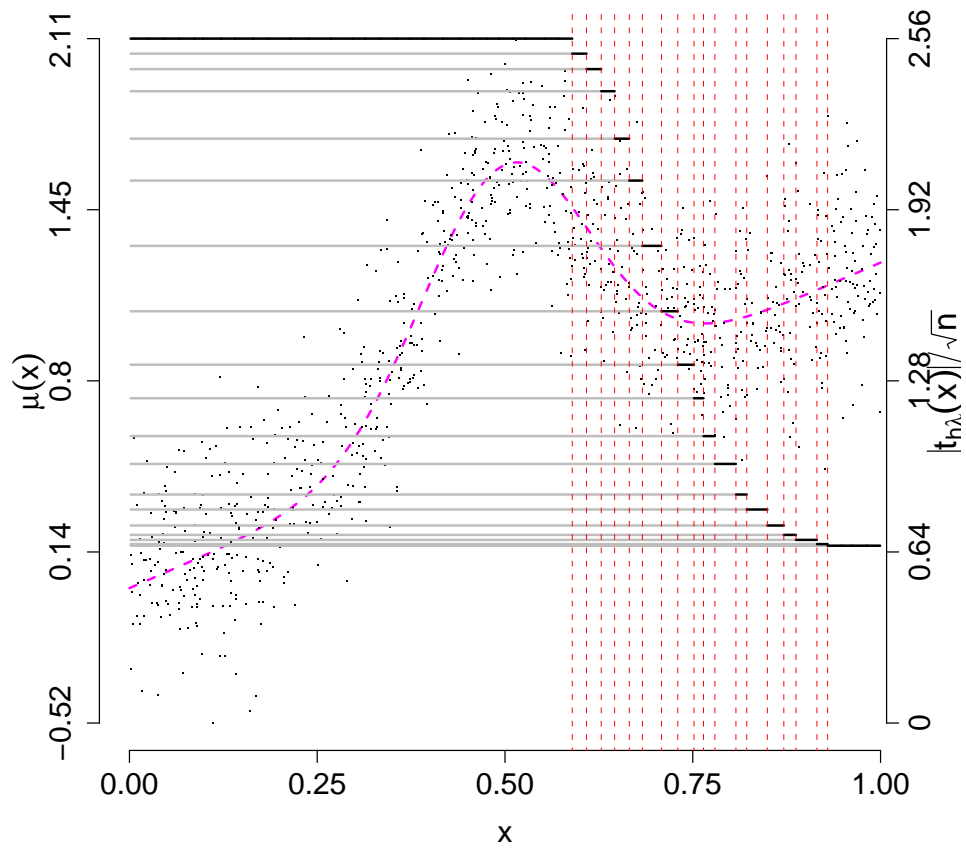


Figure 7: Plot of bandwidth chosen by maximizing the absolute  $t$ -statistic for all  $x_0$ , using one simulated data set, for the case  $n = 1000$ , and  $\sigma^2 = 0.05$  when using the sloped bump model.

weighting function (the `loess()` function implemented in R). The smoothing parameter, called `span`, is chosen by minimizing the generalized cross validation; the chosen value of 0.96 means that the local neighborhood is chosen sufficiently large to include 96% of the data.

Figure 10 shows the estimated correlation curve,  $\hat{\rho}_{h,\lambda}(\cdot)$ , as described in Remark 2.1. Again, the windows required for the local slope and variance estimates are chosen to maximize the absolute  $t$ -statistics.

This problem is well-suited to our approach. We seek those features of the bivariate relationship which these data have the most power to reveal. In Figures 8 and 10, we observe that there are two major features, one representing when there is a decrease in the exchange rate relative to the previous day (return less than one), and another feature representing when there is an increase in the exchange rate relative to the previous day. This finding is consistent with the discussion in Karpoff (1987), who cites several studies which show a positive correlation between absolute return and volume.

## 4 Appendix

**Lemma 4.1:** Let  $A(x) \equiv \int_{-\infty}^x g(t) dt$  for some integrable function  $g(\cdot)$ . If  $g''(\cdot)$  is bounded and continuous in a neighborhood of  $x$ , then

$$\begin{aligned} A(x+h) - A(x-h) &= 2A'(x)h + A'''(x)h^3/3 + o(h^3) \\ &= 2g(x)h + g''(x)h^3/3 + o(h^3). \end{aligned} \tag{9}$$

**Proof:**

$$A(x_0 \pm h) = A(x_0) + A'(x_0)(\pm h) + A''(x_0)h^2/2 + A'''(x_0)(\pm h^3)/6 + o(h^3). \square$$

**Lemma 4.2:** For constants  $c_1, c_2, c_3$ , and  $c_4$  with  $c_1 \neq 0$ ,

- (a)  $(c_1 + c_2h^2)^{-1} = c_1^{-1} - c_2c_1^{-2}h^2 + o(h^2)$  and
- (b)  $(c_1 + c_2h^2)^{-1}(c_3 + c_4h^2) = c_1^{-1}(c_3 + c_4h^2) - c_2c_3c_1^{-2}h^2 + o(h^2)$ .

**Proof:** For part (a), Taylor expand  $g(t) = (c_1 + c_2t)^{-1}$  around  $t = 0$ . Part (b) follows directly from (a).  $\square$

**Lemma 4.3:** If  $f''(\cdot)$  is continuous and bounded in a neighborhood of  $x_0$ , then

- (a)  $P_h(x_0) = 2f(x_0)h + f''(x_0)h^3/3 + o(h^3)$ ,

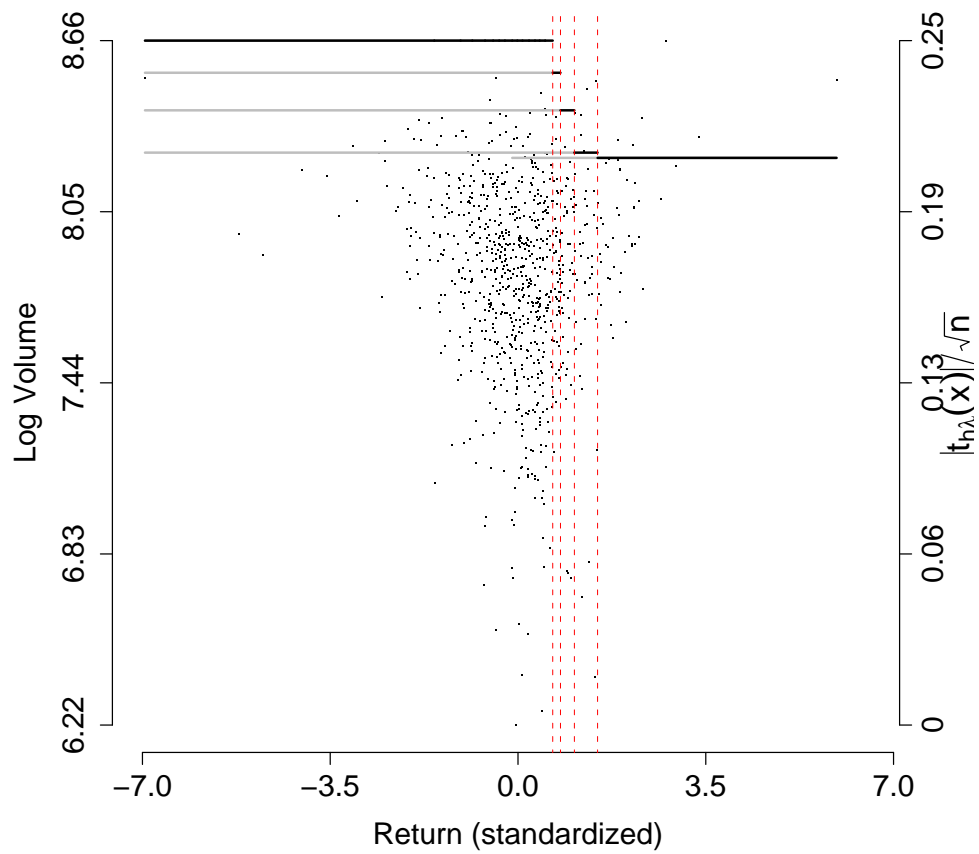


Figure 8: Results of analysis of Japanese Yen to Dollar exchange rates. Plot shows bandwidths chosen by maximizing the absolute  $t$ -statistics for all  $x_0$ .

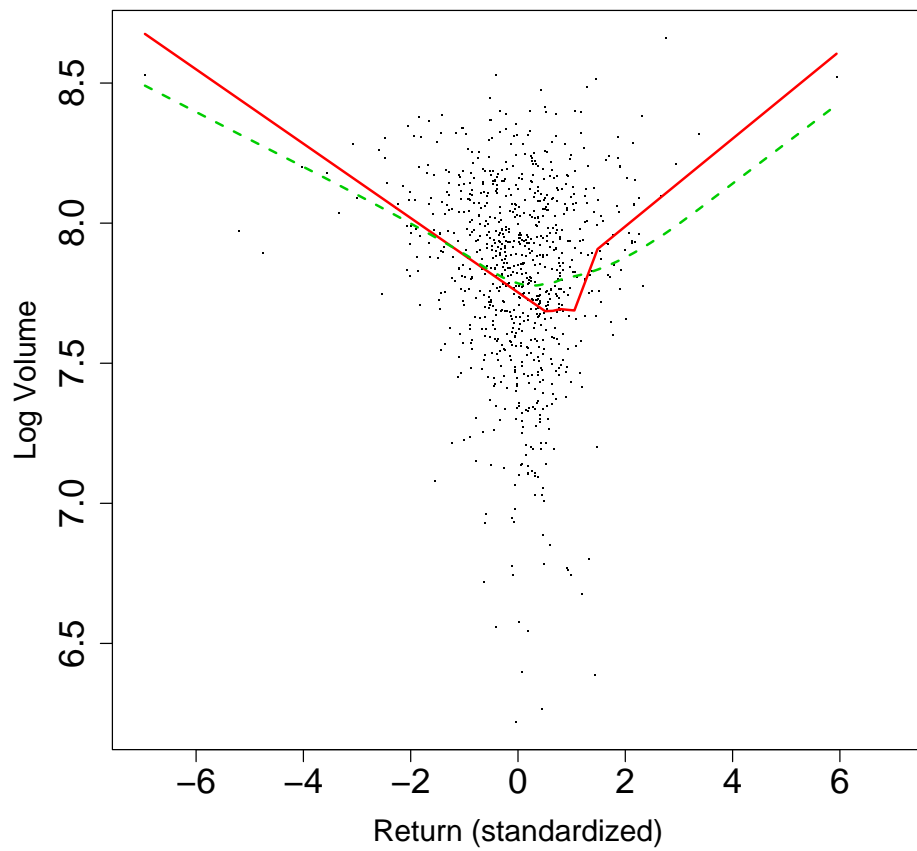


Figure 9: Results of analysis of Japanese Yen to Dollar exchange rates. Plot shows  $\hat{\mu}_{h,\lambda}(x)$  for various values of  $x$  for both bandwidth selected by maximizing the absolute  $t$ -statistics (solid), compared with fit using the R function `loess()` with smoothing parameter `span = 0.96` (dashed).

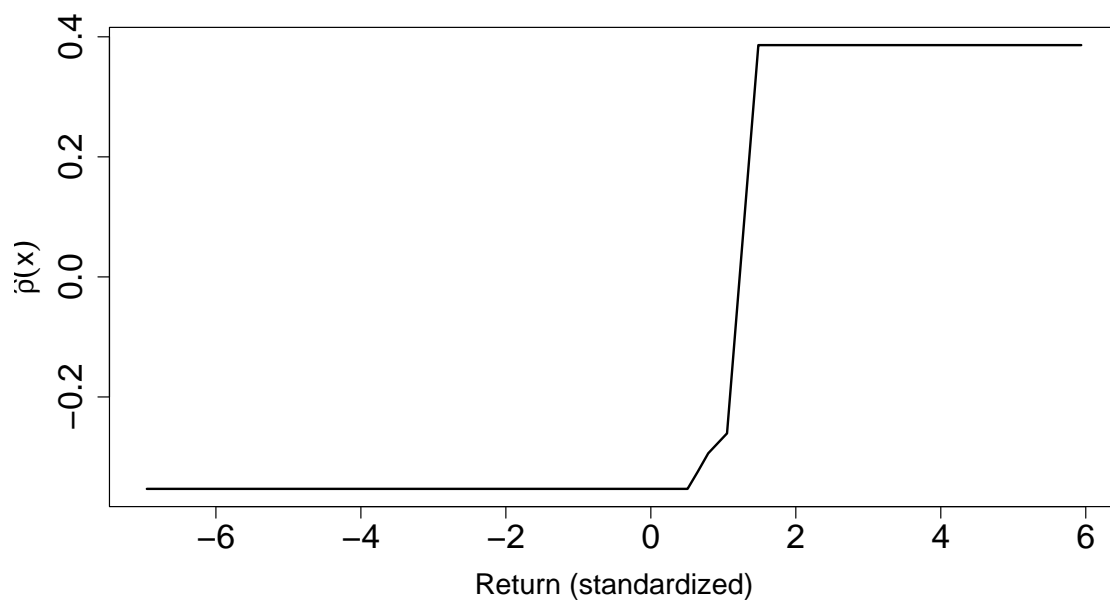


Figure 10: Results of analysis of Japanese Yen to Dollar exchange rates. Plot shows estimated correlation curve,  $\hat{\rho}_{h,\lambda}(\cdot)$  with window selected by maximizing the absolute  $t$ -statistics, as described in Remark 2.1.

$$(b) \quad E_h(X) = x_0 + [f'(x_0)/f(x_0)] h^2/3 + o(h^2),$$

$$(c) \quad \text{Var}_h(X) = h^2/3 + o(h^2),$$

$$(d) \quad [P_h(x_0)/\text{Var}_h(X)] = 6f(x_0) h^{-1} + O(1), \text{ and}$$

$$(e) \quad \begin{aligned} \text{Cov}_h(X, Y) &= \frac{\gamma\theta}{2hf(x_0)} \left\{ -\frac{1}{3}m_0(W) f'(x_0) h^2 + O(\theta h^2) \right. \\ &\quad \left. + f(x_0) m_1(W) \theta + f'(x_0) m_2(W) \theta^2 + o(h^2) + o(\theta^2) \right\}. \end{aligned}$$

**Proof:**

**Part (a):** Use Equation (9) with  $g(\cdot) = f(\cdot)$ .

**Part (b):** Set  $Z \equiv X - x_0$ , then  $E_h(X) = x_0 + E_h(Z)$  with  $f_Z(z) = f(z + x_0)$ . Write  $f_0(z) = f_Z(z)$ , then

$$P_h(x_0) E_h(Z) = \int_{-h}^h z f_0(z) dz = A(h) - A(-h).$$

Now Equation (9) with  $g(z) = z f_0(z)$ ,  $g(0) = 0$ ,  $g''(0) = 2f'(x_0)$  implies that

$$E_h(Z) = [2f'(x_0) h^3/3 + o(h^3)] / P_h(x_0) = (f'(x_0)/f(x_0)) h^2/3 + o(h^2)$$

by Lemma 4.2. The result follows.

**Part (c):** Set  $Z \equiv X - x_0$ , then  $\text{Var}(X|N_h(x_0)) = \text{Var}(Z|N_h(0))$  and  $f_Z(0) = f(x_0)$ . Write  $f_0(z)$  for  $f_Z(z)$ . We have

$$E_h(Z^2) = E(Z^2|N_h(x_0)) = \frac{1}{P_h(x_0)} \int_{-h}^h z^2 f_0(z) dz = \frac{[A(h) - A(-h)]}{P_h(x_0)}.$$

Now use Lemma 4.1 with  $g(z) = z^2 f_0(z)$ . We have  $g(0) = g'(0) = 0$ ,  $g''(0) = 2f_0(0) = 2f(x_0)$ ,

$$E_h(Z^2) = \frac{2f(x_0) h^2/3}{2f(x_0) + f''(x_0) h^2/3} = h^2/3 + o(h^2),$$

$$\text{Var}_h(Z) = E_h(Z^2) - [E_h(Z)]^2 = h^2/3 + o(h^2).$$

**Part (d):**

$$[P_h(x_0)/\text{Var}_h(X)] = \frac{2f(x_0) h + f''(x_0) h^3/3 + o(h^3)}{h^2/3 + o(h^2)} = 6f(x_0) h^{-1} + O(1).$$

**Part (e):** Begin with the expression for  $\text{Cov}_h(X, Y)$  given in Equation (3) and substitute in  $P_h(x_0)$  and  $E_h(X)$  given in parts (a) and (b) of this lemma. This gives the following:

$$\begin{aligned} \text{Cov}_h(X, Y) &= \gamma\theta (2f(x_0)h + f''(x_0)h^3/3 + o(h^3))^{-1} \\ &\times \left[ \int_{-1}^1 s\theta W(s) f(x_0 + s\theta) ds \right. \\ &\left. - \int_{-1}^1 \frac{f'(x_0)h^2}{f(x_0)3} W(s) f(x_0 + s\theta) ds + o(h^2) \right]. \end{aligned}$$

Then using the Taylor expansion of  $f(\cdot)$  around  $x_0$  we see that

$$\begin{aligned} \int_{-1}^1 s\theta W(s) f(x_0 + s\theta) ds &= f(x_0)\theta m_1(W) + f'(x_0)\theta^2 m_2(W) + o(\theta^2) \quad \text{and} \\ \int_{-1}^1 \frac{f'(x_0)h^2}{f(x_0)3} W(s) f(x_0 + s) ds &= m_0(W) f'(x_0)h^2/3 + o(h^2) + o(\theta^2) \end{aligned}$$

where the terms of this second equation which contain  $h^2\theta$  or  $h^2\theta^2$  are  $o(h^2)$  since we are considering behavior under both  $\theta$  and  $h$  going to zero. Combining, we see that

$$\begin{aligned} \text{Cov}_h(X, Y) &= (2f(x_0)h + o(h^2))^{-1} \\ &\times (f(x_0)\theta m_1(W) + f'(x_0)m_2(W)\theta^2 \\ &- m_0(W) f'(x_0)h^2/3 + o(h^2) + o(\theta^2)), \end{aligned}$$

and the result follows immediately.  $\square$

**Lemma 4.4:** Let  $A(x) \equiv \int_{-\infty}^x g(t) dt$  for some integrable function  $g(\cdot)$ . Then

$$\begin{aligned} D &\equiv A(x + (1 + \lambda)h) - A(x - (1 - \lambda)h) \\ &= 2A'(x)h + 2A''(x)\lambda h^2 + A'''(x)(1 + 3\lambda^2)h^3/3 + o(h^3) \\ &= 2g(x)h + 2g'(x)\lambda h^2 + g''(x)(1 + 3\lambda^2)h^3/3 + o(h^3) \end{aligned}$$

provided  $g''(\cdot)$  is bounded and continuous in a neighborhood of  $x$ .

**Proof:**

$$\begin{aligned} A(x + (1 + \lambda)h) &= A(x) + A'(x)(1 + \lambda)h + A''(x)(1 + \lambda)^2 h^2/2 \\ &+ A'''(x)(1 + \lambda)^3 h^3/6 + o(h^3). \end{aligned}$$

$$\begin{aligned} A(x - (1 - \lambda)h) &= A(x) - A'(x)(1 - \lambda)h + A''(x)(1 - \lambda)^2 h^2/2 \\ &- A'''(x)(1 - \lambda)^3 h^3/6 + o(h^3). \quad \square \end{aligned}$$

**Lemma 4.5:** If  $f''(\cdot)$  is continuous and bounded in a neighborhood of  $x_0$ , then

- (a)  $P_{h,\lambda}(x_0) = 2f(x_0)h + 2f'(x_0)\lambda h^2 + o(h^2)$ ,
- (b)  $E_{h,\lambda}(X) = x_0 + \lambda h + [f'(x_0)/f(x_0)]h^2/3 + o(h^2)$ ,
- (c)  $\text{Var}_{h,\lambda}(X) = h^2/3 + o(h^2)$ , and
- (d)  $[P_{h,\lambda}(x_0)/\text{Var}_{h,\lambda}(X)] = 6f(x_0)h^{-1} + O(1)$ .

**Proof:**

**Part (a):** Use Lemma 4.4 with  $g(\cdot) = f(\cdot)$ .

**Part (b):** Again set  $Z \equiv X - x_0$ , then

$$\begin{aligned} E_{h,\lambda}(Z) &= (2f(x_0)\lambda h^2 + 2f'(x_0)/3(1 + 3\lambda^2)h^3 + o(h^3)) / (P_{h,\lambda}(x_0)) \\ &= \frac{2f(x_0)\lambda h + 2f'(x_0)(1 + 3\lambda^2)h^2/3 + o(h^2)}{2f(x_0) + 2f'(x_0)\lambda h + o(h)} \\ &= \lambda h + [f'(x_0)/f(x_0)](1 + 3\lambda^2)h^2/3 - [f'(x_0)/f(x_0)]\lambda^2 h^2 \\ &= \lambda h + [f'(x_0)/f(x_0)]h^2/3 + o(h^2). \end{aligned}$$

**Part (c):**

$$\begin{aligned} E_{h,\lambda}(Z^2) &= \frac{1}{P_{h,\lambda}(x_0)} [2f(x_0)(1 + 3\lambda^2)h^3/3] \\ &= (1 + 3\lambda^2)h^2/3 = h^2/3 + \lambda^2 h^2 + o(h^2). \end{aligned}$$

$$\text{Var}_{h,\lambda}(Z) = h^2/3 + \lambda^2 h^2 - \lambda^2 h^2 + o(h^2) = h^2/3 + o(h^2).$$

**Part (d):**

$$\frac{P_{h,\lambda}(x_0)}{\text{Var}_{h,\lambda}(X)} = \frac{2f(x_0)h + 2f'(x_0)\lambda h^2 + o(h^2)}{h^2/3 + o(h^2)} = 6f(x_0)h^{-1} + O(1). \quad \square$$

## Acknowledgments

We are grateful to Alex Samarov for many helpful comments.

## References

- AIT-SAHALIA, Y., BICKEL, P. J. and STOKER, T. M. (2001). Goodness-of-fit Tests for Kernel Regression with an Application to Option Implied Volatilities. *J. of Econ.*, **105** 363–412.

- AZZALINI, A., BOWMAN, A. W. and HARDLE, W. (1989). On the Use of Nonparametric Regression for Model Checking. *Biometrika*, **76** 1–11.
- BEHNEN, K. and HUŠKOVÁ, M. (1984). A simple algorithm for the adaptation of scores and power behavior of the corresponding rank test. *Comm. in Stat. - Theory and Methods*, **13** 305–325.
- BEHNEN, K. and NEUHAUS, G. (1989). *Rank Tests with Estimated Scores and their Applications*. Teubner, Stuttgart.
- BELL, C. B. and DOKSUM, K. A. (1966). “Optimal” One-Sample Distribution-Free Tests and Their Two-Sample Extensions. *Ann. Math. Stat.*, **37** 120–132.
- BICKEL, P. J. and DOKSUM, K. A. (2001). *Mathematical Statistics. Basic Ideas and Selected Topics, Volume I*. Prentice Hall, New Jersey.
- BICKEL, P. J., RITOV, Y. and STOKER, T. M. (2006). Tailor-made Tests for Goodness-of-Fit to Semiparametric Hypotheses. *Ann. Statist.*, **34**. To Appear.
- BJERVE, S. and DOKSUM, K. (1993). Correlation Curves: Measures of Association as Functions of Covariate Values. *Ann. Statist.*, **21** 890–902.
- BLYTH, S. (1993). Optimal Kernel Weights Under a Power Criterion. *J. Amer. Statist. Assoc.*, **88** 1284–1286.
- CHAUDHURI, P. and MARRON, J. S. (1999). SiZer for Exploration of Structures in Curves. *J. Amer. Statist. Assoc.*, **94** 807–823.
- CHAUDHURI, P. and MARRON, J. S. (2000). Scale Space View of Curve Estimation. *Ann. Statist.*, **28** 408–428.
- CLAESKENS, G. and HJORTH, N. L. (2003). The Focused Information Criterion. *J. Amer. Statist. Assoc.*, **98** 900–916.
- DOKSUM, K. (1966). Asymptotically Minimax Distribution-free Procedures. *Ann. Math. Stat.*, **37** 619–628.
- DOKSUM, K., BLYTH, S., BRADLOW, E., MENG, X.-L. and ZHAO, H. (1994). Correlation Curves as Local Measures of Variance Explained by Regression. *J. Amer. Statist. Assoc.*, **89** 571–582.

- DOKSUM, K. and SAMAROV, A. (1995). Nonparametric Estimation of Global Functionals and a Measure of the Explanatory Power of Covariates in Regression. *Ann. Statist.*, **23** 1443–1473.
- DOKSUM, K. A. and FRODA, S. (2000). Neighborhood Correlation. *J. of Statist. Planning and Inference*, **91** 267–294.
- DONOHO, D. L. and LIU, R. C. (1991a). Geometrizing Rates of Convergence, II. *Ann. Statist.*, **19** 633–667.
- DONOHO, D. L. and LIU, R. C. (1991b). Geometrizing Rates of Convergence, III. *Ann. Statist.*, **19** 668–701.
- EINMAHL, V. and MASON, D. (2005). Uniform in Bandwidth Consistency of Kernel-type Function Estimators. *Ann. Statist.*, **33** 1380–1403.
- FAN, J. (1992). Design-adaptive Nonparametric Regression. *J. Amer. Statist. Assoc.*, **87** 998–1004.
- FAN, J. (1993). Local Linear Regression Smoothers and Their Minimax Efficiencies. *Ann. Statist.*, **21** 196–216.
- FAN, J., ZHANG, C. and ZHANG, J. (2001). Generalized Likelihood Ratio Statistics and Wilks Phenomenon. *Ann. Statist.*, **29** 153–193.
- FAN, J. Q. and GIJBELS, I. (1996). *Local Polynomial Modelling and Its Applications*. Chapman & Hall, London.
- GASSER, T., MULLER, H.-G. and MAMMITZSCH, V. (1985). Kernels for Nonparametric Curve Estimation. *J. Roy. Stat. Soc., Ser. B*, **47** 238–252.
- GODTLIEBSEN, F., MARRON, J. and CHAUDHURI, P. (2004). Statistical Significance of Features in Digital Images. *Image and Vision Computing*, **22** 1093–1104.
- HAJEK, J. (1962). Asymptotically Most Powerful Rank-Order Tests. *Ann. Math. Stat.*, **33** 1124–1147.
- HALL, P. and HART, J. D. (1990). Bootstrap Test for Difference Between Means in Nonparametric Regression. *J. Amer. Statist. Assoc.*, **85** 1039–1049.
- HALL, P. and HECKMAN, N. E. (2000). Testing for Monotonicity of a Regression Mean by Calibrating for Linear Functions. *Ann. Statist.*, **28** 20–39.

- HART, J. (1997). *Nonparametric Smoothing and Lack-of-Fit Tests*. Springer-Verlag, New York.
- INGSTER, Y. I. (1982). Minimax nonparametric detection of signals in white Gaussian noise. *Problems in Information Transmission*, **18** 130–140.
- KARPOFF, J. M. (1987). The Relation Between Price Changes and Trading Volume: A Survey. *J. of Fin. and Quant. Anal.*, **22** 109–126.
- KENDALL, M. and STUART, A. (1961). *The Advanced Theory of Statistics, Volume II*. Charles Griffin & Company, London.
- LEHMANN, E. (1999). *Elements of Large Sample Theory*. Springer, New York.
- LEPSKI, O. V. and SPOKOINY, V. G. (1999). Minimax Nonparametric Hypothesis Testing: The Case of an Inhomogeneous Alternative. *Bernoulli*, **5** 333–358.
- LINHART, H. and ZUCCHINI, W. (1986). *Model Selection*. John Wiley & Sons, New York.
- NEYMAN, J. (1959). Optimal Asymptotic Tests of Composite Hypotheses. In *Probability and Statistics: The Harold Cramer Volume* (U. Greenlander, ed.). John Wiley & Sons, New York, 213–234.
- RAZ, J. (1990). Testing for No Effect When Estimating a Smooth Function by Nonparametric Regression: A Randomization Approach. *J. Amer. Statist. Assoc.*, **85** 132–138.
- SEN, P. K. (1996). Regression Rank Scores Estimation in ANOCOVA. *Ann. Statist.*, **24** 1586–1601.
- STUTE, W. and ZHU, L. (2005). Nonparametric Checks For Single-Index Models. *Ann. Statist.*, **33** 1048–1083.
- ZHANG, C. M. (2003a). Adaptive Tests of Regression Functions via Multi-scale Generalized Likelihood Ratios. *Canadian J. Statist.*, **31** 151–171.
- ZHANG, C. M. (2003b). Calibrating the Degrees of Freedom for Automatic Data Smoothing and Effective Curve Checking. *J. Amer. Statist. Assoc.*, **98** 609–628.