EDGEWORTH APPROXIMATIONS FOR SEMIPARAMETRIC INSTRUMENTAL VARIABLE ESTIMATORS AND TEST STATISTICS*

by

Oliver Linton
London School of Economics and Political Science

Contents:

The Suntory Centre
Suntory and Toyota International Centres for Economics and Related Disciplines
London School of Economics and Political Science
Houghton Street
London WC2A 2AE
Tel.: 020-7405 7686

Discussion Paper
No.EM/00/399
July 2000

* I would like to thank Zhiejie Xiao and Moto Shintani for excellent research assistance, Joel Horowitz, Yuichi Kitamura, Peter Phillips, and Tom Rothenberg for helpful comments, and Germán Aneiros for pointing out a mistake in an earlier draft of this paper. I would also like to thank three referees for extensive comments which greatly improved the paper. I am grateful to the National Science Foundation for financial support.

Abstract

We establish the validity of higher order asymptotic expansions to the distribution of a version of the nonlinear semiparametric instrumental variable considered in Newey (1990) as well as to the distribution of a Wald statistic derived from it. We employ local polynomial smoothing with variable bandwidth, which includes local linear, kernel, and [a version of] nearest neighbour estimates as special cases. Our expansions are valid to order $n^{-2\epsilon}$ for some $0 < \epsilon < ½$, where ϵ depends on the smoothness and dimensionality of the data distribution and on the order of the polynomial chosen by the practitioner. We use the expansions to define optimal bandwidth selection methods for both estimation and testing problems and apply our methods to simulated data.

Keywords: Bandwidth selection; Edgeworth approximation; instrumental variables; kernel estimation; local polynomials.

JEL No.: C14

1 Introduction

Instrumental variables and the related Generalized Method of Moments estimation procedures are widely used and taught in econometrics courses. In recent years these methods have been viewed as being semiparametric, in the sense that the joint distribution of the data is unspecified apart from a finite number of moment conditions. Frequently, this information is in the form of conditional moments, which implies, generally, an infinite number of unconditional moment conditions, although a certain finite dimensional combination of them gives full efficiency – see Chamberlain (1987) and Newey (1986, 1988, 1990). The efficient instrument function involves an unknown conditional expectation. Therefore, to fully exploit the given moment conditions, it is necessary to use nonparametric regression techniques to estimate the optimal instruments. Newey (1990) established the asymptotic properties of a semiparametric instrumental variable estimator $\widehat{\theta}$ based on a nonparametric estimate $\widehat{g}$ (specifically, nearest neighbors and series estimates) of the optimal instrument function g. Under regularity conditions, he showed that $n^{1/2}(\widehat{\theta} - \theta_0)$ is asymptotically normal with zero mean and asymptotic variance ω, where n is sample size and ω is a positive definite matrix; in fact, $\widehat{\theta}$ is asymptotically equivalent to the procedure based on the true unknown optimal instrument function. There are many applications of this estimation procedure in sample selection and binary choice models as well as other microeconometric contexts.

We have argued elsewhere, Linton (1995, 1996a,b), that the first-order asymptotics of semiparametric procedures can be misleading and unhelpful.[1] The limiting variance matrix ω does not depend on the specific details of how $\widehat{g}$ is constructed, and thus sheds no light on how to implement this important part of the procedure. Specifically, bandwidth choice cannot be addressed by using the first-order theory alone. Also, the relative merits of alternative first-order equivalent implementations, e.g., one-step procedures, cannot be determined by the first-order theory alone. Finally, to show when bootstrap methods can provide asymptotic refinements for asymptotically pivotal statistics requires some knowledge of higher-order properties – see Horowitz (1995). This motivates the study of higher-order expansions. Carroll and Härdle (1989) was to our knowledge the first published paper that developed second-order mean squared error expansions for a semiparametric, i.e., smoothing-based but root-n consistent, procedure, in the context of a heteroskedastic linear regression. Härdle, Hart, Marron, and Tsybakov (1992) developed expansions for scalar average derivatives which was extended to the multivariate case, actually only the simpler situation of density-weighted average derivatives, by Härdle and Tsybakov (1993); these papers used the expansions to develop automatic bandwidth selection routines. This work was extended to the slightly more general case of density-weighted averages by Powell and Stoker (1996). Linton (1995, 1996a) developed similar expansions

[1] See also the monte carlo evidence presented in for example Hsieh and Manski (1987).

for the partially linear model and the heteroskedastic linear regression model and provided some results on the optimality of the bandwidth selection procedures proposed therein. Xiao and Phillips (1996) worked out the same approximations for a time series regression model with serial correlation of unknown form; Xiao and Linton (1997) give the analysis for Bickel's (1982) adaptive estimator in the linear regression model; Linton and Xiao (1997) works out the approximations for the nonlinear least squares and profile likelihood estimators in a semiparametric binary choice model. Robinson (1995) revisited the average derivative estimation problem, in particular the density-weighted version of this procedure, which is easier to handle, and proved the validity of the Berry-Esséen theorem for this estimator. This work was recently extended by Nishiyama and Robinson (1997) to include a proof of the validity of an Edgeworth expansion for the same estimator. These last two works have contributed greatly to the rigour of the analysis, albeit in a simple setting. They also point out that under certain circumstances the dominant correction effect on the distribution of the estimator is of order $n^{-1/2}$, as in parametric situations, and unrelated to the smoothing operation itself. This point has also been made by Liang and Cheng (1993) in their study of the partially linear regression model; in fact, they take the analysis one step further and argue that the order $n^{-1/2}$ term is optimal in a certain sense. See also Liang (1995) for similar results from the point of view of Bahadur efficiency. Some other important developments include work by Horowitz (1998), who investigates bootstrap in non- and semiparametric models with a view to determining bandwidth and providing asymptotic refinements.

In this paper, we develop second-order approximations for an implicitly defined semiparametric instrumental variable estimator similar to that considered by Newey (1990) except that we also weight for heteroskedasticity of unknown form. Previous work by the author (Linton, 1995, 1996a) developed asymptotic expansions for the cumulants of standardized semiparametric estimators in the heteroskedastic linear regression and the partially linear model. Specifically, the asymptotic mean squared error is of the form $\omega + n^{-2\epsilon}\omega_c$, where ω_c is a positive definite matrix and $0 < \epsilon < 1/2$. The correction term is of larger magnitude than in parametric procedures and reflects the method used to estimate g and the smoothness of this function. These estimators are both explicitly defined. Furthermore, a fixed design assumption was maintained. In this paper we derive the second-order properties of the implicitly defined instrumental variable estimator in the more primitive random design. We calculate approximations to the first four cumulants [valid to order $n^{-2\epsilon}$] and prove that the formal Edgeworth approximation based on them provides a valid approximation to the distribution of the standardized estimator correct to the same order of magnitude. As in Nishiyama and Robinson (1997), the order $n^{-1/2}$ term, due to bias and skewnesses, can dominate in the distributional approximation, while the smoothing-based terms affect only the variance, which is, under some restrictions, of a smaller magnitude. We use the distributional approximation to

define an optimal bandwidth for a general class of criteria, and develop practical bandwidth selection methods. We also examine a Wald test statistic of general nonlinear restrictions, providing the Edgeworth approximation to its distribution under the null hypothesis. The second-order properties of the test depend on which standard error and which bandwidths are used. By using less smoothing when estimating the standard error matrix estimator, we obtain better asymptotic performance in terms of null rejection frequency. In this case, the size distortion [the difference between the nominal and actual null rejection frequency] of the test is of the same magnitude, i.e., order $n^{-2\epsilon}$, as the variance of the estimator. We define an optimal bandwidth in this context as one that minimizes the second-order size distortion and then use our expansions to suggest a feasible bandwidth selection method based on this notion. The optimal bandwidth is of a similar form to that derived in the estimation problem and does not to depend on the level of the test.

This paper is organized as follows. In section 2 we define the model of interest and introduce the estimators. In section 3 we derive the higher-order asymptotic properties of the estimator. In section 4 we examine standard errors and a Wald statistic. In section 5 we discuss optimality and bandwidth choice for estimation. In section 6 we present some simulation results. All derivations are given in the appendix.

For any vectors $\mathbf{x} = (x_1, \ldots, x_d)'$ and $\mathbf{a} = (a_1, \ldots, a_d)'$, define $|\mathbf{x}| = \sum_{j=1}^d x_j$, $\mathbf{x}! = x_1! \times \cdots \times x_d!$, and $\mathbf{x}^{\mathbf{a}} = (x_1^{a_1}, \ldots, x_d^{a_d})'$, also let

$$\partial^{\mathbf{a}} g(\mathbf{x}) = \frac{\partial^{|\mathbf{a}|} g}{\partial x_1^{a_1} \cdots \partial x_d^{a_d}}(\mathbf{x}).$$

Let $\Phi_{\mu,V}$ [$\phi_{\mu,V}$] and $F_{p,\infty}$ [$f_{p,\infty}$] be the distribution functions [densities] of a $N(\mu, V)$ and a χ^2_p random variable respectively, and let $\|A\| = \mathrm{tr}^{1/2}(A'A)$ be the Euclidean norm of any $p \times m$ matrix A.

2 Estimator and Assumptions

Suppose that there is a population random variable $Z = (Y, X)$ and that there is an independent and identically distributed sample $\{Z_i\}_{i=1}^n$ drawn from this population. We assume that there is a unique $\theta_0 \in \Theta \subseteq \mathbb{R}^p$ satisfying the conditional moment conditions: $E\{\rho(Z, \theta_0) \,|X\} = 0$ with probability one, where $\rho(z, \theta)$ is an m by 1 vector of functions. This implies the unconditional moment conditions

$$E\left\{A(X)\rho(Z, \theta_0)\right\} = 0, \tag{1}$$

for any $p \times m$ matrix $A(X)$ [for which the expectation exists]. The sample version of (1) is the basis of estimation as described in many previous papers, see for example Amemiya (1974). Suppose that with probability one $E[\rho(Z, \theta_0)\rho(Z, \theta_0)' \,|X] = \Omega(X)$ for some unknown function Ω, then the

optimal weighting matrix is proportional to $D(X_i)\Omega(X)^{-1}$, where $D(X_i) = D(X_i;\theta_0)$ with $D(X_i;\theta) = E\{\partial\rho(Z_i,\theta)/\partial\theta' \,|X_i\}$, in which case the asymptotic variance of the standardized estimator is J^{-1} with $J = E\{D(X)\Omega(X)^{-1}D(X)'\}$.

We shall suppose that the optimal instrument function $D(\cdot)$ and the variance function $\Omega(\cdot)$ are smooth but otherwise of unknown form. Our estimation strategy is similar to Newey (1990) except that we estimate both the instrument function and the heteroskedasticity nonparametrically. We first obtain preliminary root-n consistent estimates $\widetilde{\theta}$ of θ, which are obtained as any solution to $\sum_{i=1}^n A(X_i)\rho(Z_i,\widetilde{\theta}) = 0$ for any known function $A(\cdot)$. We then consider estimators $\widehat{\theta}$ that satisfy

$$\widehat{s}(\widehat{\theta}) = \frac{1}{n}\sum_{i=1}^n \widetilde{D}(X_i;\widetilde{\theta})\widetilde{\Omega}(X_i)^{-1}\rho(Z_i,\widehat{\theta}) = 0, \tag{2}$$

where:

$$\widetilde{D}(X_i;\widetilde{\theta}) = \sum_{j\neq i} w_{ij}\frac{\partial\rho(Z_j,\widetilde{\theta})}{\partial\theta'} \tag{3}$$

$$\widetilde{\Omega}(X_i;\widetilde{\theta}) = \sum_{j\neq i} w_{ij}\rho(Z_j,\widetilde{\theta})\rho(Z_j,\widetilde{\theta})' \tag{4}$$

are nonparametric estimates of the instrument function and conditional covariance matrix. Here, w_{ij} are nonparametric smoother weights. Our results will be proven for a general class of variable bandwidth local polynomial weights, which we now introduce using the definitions of Masry (1996ab). For a scalar dependent variable $\{T_i\}_{i=1}^n$, let the parameter vector $\widehat{\boldsymbol{\alpha}}(X_i)$ minimize the criterion function

$$\sum_{j\neq i} K\left(\frac{X_j - X_i}{h_{ni}}\right)\{T_j - \sum_{\mathbf{a}:0\leq|\mathbf{a}|\leq q-1} \boldsymbol{\alpha}_{\mathbf{a}}\cdot(X_j - X_i)^{\mathbf{a}}\}^2, \tag{5}$$

with respect to $\boldsymbol{\alpha}$, where $\boldsymbol{\alpha}$ [and hence $\widehat{\boldsymbol{\alpha}}(X_i)$] consists of all scalars $\boldsymbol{\alpha}_{\mathbf{a}}$ [$\widehat{\boldsymbol{\alpha}}_{\mathbf{a}}(X_i)$] indexed by the vector $\mathbf{a}$ which runs through all possibilities with $0 \leq |\mathbf{a}| \leq q-1$. Here, $K(\cdot)$ is a multivariate kernel function and h_{ni} is a bandwidth sequence that is allowed to vary with the evaluation points and satisfies regularity conditions specified below. The minimizing value $\widehat{\boldsymbol{\alpha}}(X_i)$ is linear in T_j; in particular $\widehat{\alpha}_0(X_i)$, which is the estimate of $E(T_i|X_i)$, satisfies $\widehat{\alpha}_0(X_i) = \sum_{j\neq i} w_{ij}T_j$ for some weights w_{ij} that depend on only the independent variables. These are the weights that we use in (3). Our class of smoothers includes local constant kernel [by taking $h_{ni} = h_n$ and $q = 1$] and (approximate) nearest neighbor [by taking h_{ni} inversely proportional to the covariate density and $q = 1$, see Härdle and Linton (1994)], as well as local linear and higher order polynomials. The odd order polynomial estimators with constant bandwidth have been extensively praised for their ability to adapt to the

design density and to the effective boundary region, see the recent book of Fan and Gijbels (1996) for discussion.

Let $\rho_i = \rho(Z_i, \theta_0)$, $\eta_i = \partial\rho(Z_i, \theta_0)/\partial\theta' - D(X_i)$, and $\zeta_i = \rho(Z_i, \theta_0)\rho(Z_i, \theta_0)' - \Omega(X_i)$, and define also $\eta_i^\dagger = \eta_i - D(X_i)\Omega(X_i)^{-1}\zeta_i$. We make the following assumptions about the sampling scheme and smoothing weights:

ASSUMPTION A1. *The distribution of the d-dimensional vector X has compact support $\mathcal{S} = \cap\{F$: F closed, $\Pr(X \in \mathbb{R}^d \backslash F) = 0\}$, and is absolutely continuous with respect to Lebesgue measure restricted to $\mathcal{S}$. It has density f, which satisfies $\inf_x f(x) \geq \underline{f} > 0$. We also suppose that for all $\epsilon > 0$, $0 < \underline{c} \leq \mu(\mathcal{S} \cap (\partial\mathcal{S})^\epsilon)/\epsilon^d \leq \overline{c} < \infty$, where μ is Lebesgue measure and $(\partial\mathcal{S})^\epsilon = \{x : \|x - y\| < \epsilon$ for some $y \in \partial\mathcal{S}\}$.*

ASSUMPTION A2. (a) *θ_0 is an interior element of the compact set $\Theta \subseteq \mathbb{R}^p$;* (b) *$\rho(Z, \theta)$ is 4-times continuously differentiable in θ for each $\theta \in \Theta$ with probability one;* (c) *for any $\{\ell_j\}_{j=1}^p$ with $\ell_1 + \ldots + \ell_p = \ell$ with $\ell = 0, 1, 2, 3, 4$, we have $E[\left|\partial^\ell \rho(Z_i, \theta_0)/\partial\theta_1^{\ell_1} \cdots \partial\theta_p^{\ell_p}\right|^\nu] < \infty$ for $\nu = 1, 2, \ldots$;* (d) *$\sup_{\theta\in\Theta} \left|\partial^\ell \rho(z, \theta)/\partial\theta_1^{\ell_1} \cdots \partial\theta_p^{\ell_p}\right| \leq M_\ell(z)$ for $M_\ell(z)$ such that $E[M_\ell(Z)^\nu] < \infty$ for $\nu = 1, 2, \ldots$ and $\ell = 4$.*

ASSUMPTION A3. *The marginal density $f(\cdot)$, the regression function $D(\cdot; \theta)$, the covariance matrix $\Omega(\cdot; \theta)$, and their partial derivatives in θ at $\theta = \theta_0$ through fourth order are all r-times uniformly boundedly continuously differentiable on the interior $\mathcal{S}_0$ of $\mathcal{S}$, where $r \geq (q + 2)$. Furthermore, all elements of the $p \times p$ matrix functions $\Omega(x) = E(\rho_j \rho_j' \,|X_j = x)$, $S_1(x) = E(\eta_i^\dagger \Omega^{-1}(X_i)\eta_i^{\dagger\prime} \,|X_i = x)$ and $S_2(x_1, x_2) = E(\eta_i^\dagger \Omega^{-1}(X_i)\rho_j \rho_j' \Omega^{-1}(X_i)\eta_i^{\dagger\prime} \,|X_i = x_1, X_j = x_2)$ are continuous throughout $\mathcal{S}$ and $\mathcal{S} \times \mathcal{S}$, and $\inf_x \lambda_{\min}(\Omega(x)) \geq \underline{\lambda} > 0$.*

ASSUMPTION A4. *The kernel $K(\cdot)$ has bounded support, is symmetric about zero, and is Lipschitz continuous, i.e., there exists a constant C such that $|K(u) - K(v)| \leq C\,|u - v|$ for all u, v. Define $\|K\|_2^2 = \int K(u)^2 du$.*

ASSUMPTION A5. *The bandwidth sequence $h_{ni} = h_n(X_i)$ satisfies $h_n(x) = \tau(x)n^{-1/(2q+d)}$, where $\tau(\cdot)$ is continuously differentiable and bounded away from zero and infinity on $\mathcal{S}$ except perhaps on set of Lebesgue measure zero. Let $\underline{h}(n) = \inf_{x\in\mathcal{S}} h_n(x)$ and $\overline{h}(n) = \sup_{x\in\mathcal{S}} h_n(x)$.*

Assumption A1 (the density of X has bounded support) is quite strong. If we were interested only in the first-order theory, then one could weaken this assumption considerably at the expense, for the kernel method, of using a trimming function [this is a disadvantage of the kernel method relative to the nearest neighbors used by Robinson (1987) and Newey (1990)]. This assumption is not needed in density weighted average derivative estimation [Powell and Stoker (1996), Nishiyama and Robinson (1997)], for example, but this is because the density weighting has removed the random denominator. However, for the Edgeworth expansion developed here one would have to use a trimming device for any smoothing method if an unbounded support is to be allowed, see below. Perhaps the main restriction embodied in A1 is that the density function be strictly positive on this bounded support, although of course the lower bound $\underline{f}$ can be arbitrarily close to zero. It is possible to relax A1 for our method along the lines of Härdle, Hart, Marron, and Tsybakov (1992), replacing this assumption by conditions of the form $E\left[g(X)\left/f^{\ell}(X)\right.\right] < \infty$ for positive integers $\ell \geq 2$ and various functions $g(\cdot)$ related to $D(\cdot)$. However, we expect that this will require strong restrictions on the functions g and hence D itself, which make this 'weakening' not worth pursuing. For example, in Härdle, Hart, Marron, and Tsybakov (1992) conditions (B2) and (A4) imply that $g(\cdot)$ must approach zero at the boundary, which seems a bit difficult to justify, especially when g is the conditional second moment as in their assumption B2 − apart from anything else this excludes homoskedastic constant regression.[2] Finally, the assumption on $\mu(\mathcal{S} \cap (\partial\mathcal{S})^{\epsilon})$ is really that the Hausdorff dimension of the set $\mathcal{S} \cap (\partial\mathcal{S})^{\epsilon}$ is d; this is satisfied by regular sets with nonempty interior, such as rectangular and spherical sets, see Besicovitch (1993, p 157). It is needed to ensure that the boundary region can not be isolated from the interior.

In assumption A2 we required an infinite number of moments for ρ. This is for convenience only; the precise number of moments required is large but finite [for comparison, Hall and Horowitz (1996) assumed thirty two moments], and varies from result to result and depends on the smoothness and dimensionality conditions in a rather complicated way. In any event these conditions are only sufficient and not necessary and the method can be expected to work well in the absence of such strong conditions. Assumption A3 and Assumption A4 are fairly standard for kernel-based estimation.

Assumption A5 allows the bandwidth to vary with the estimation point through the function $\tau(\cdot)$. We have chosen the magnitude of the bandwidth to be 'optimal' in the sense that taking larger or smaller bandwidth would lead to a larger asymptotic mean squared error for $\widehat{\theta}$. With this choice of bandwidth the pointwise mean squared error of the nonparametric estimate $\widetilde{D}$ is of order $n^{-2\epsilon}$ for $\epsilon = q/(2q+d)$, which is, according to Stone (1980), the optimal rate for any estimator of D. Our use

[2]For example, suppose that $\mathcal{S} = [0,1]$, and that as $x \to 0$, $f(x) = O(x^{\varrho})$ for some $\varrho \geq 3$ [their assumption A4 requires three times contiuous differentiability]. Then, we must have $g(x) = O(x^{1-(\ell-1)\varrho})$ as $x \to 0$ otherwise there is a non-removable singularity in $g(X)/f^{\ell-1}(X)$ which would prevent the above expectation from existing.

of the word optimality is somewhat weaker and only compares estimators of the form (2)-(5) with bandwidth $h_n(x) = \tau(x)n^{-\pi}$ as π varies.

Masry (1996a) established uniform consistency with rates and asymptotic normality (1996b) for the local polynomial estimators with constant bandwidth $h_{ni} = h_n$. These results readily extend to the variable bandwidth estimator under our conditions. Define

$$B_{ni} = B_n(X_i) = \sum_{j \neq i} w_{ij} D(X_j) - D(X_i)$$

the conditional [on $X_1, \ldots, X_n$] bias of $\widetilde{D}(X_i; \theta_0)$. Under our conditions, there exists a non-random continuous function $B(\cdot)$ and an increasing sequence of interior open sets $\mathcal{S}_n \subset \mathcal{S}$ with $\lim_{n\to\infty} \mathcal{S}_n = \mathcal{S}$, for which

$$\sup_{x \in \mathcal{S}_n} |n^{\epsilon} B_n(x) - B(x)| \to 0 \tag{6}$$

with probability one. For the local linear estimator of a univariate regression function m, the bias function, i.e., $B(\cdot)$, is proportional to $m''(x)$, to be compared with $m''(x) + 2m'(x)f'(x)/f(x)$ for the Nadaraya-Watson estimator. The bias function of the nearest-neighbor estimator is proportional to

$$\frac{m''(x) + 2m'(x)f'(x)/f(x)}{8f^2(x)}, \tag{7}$$

see Härdle and Linton (1994, Theorem 3); (7) tends to be large in regions where the marginal density is small, which might explain the simulation results reported in Newey (1990). In fact, when the marginal density is normal, $E[B^2(X)]$ [for the nearest neighbor estimate] can be finite only if there are tail conditions on m''.

Finally, we shall make an assumption about the preliminary estimator $\widetilde{\theta}$ and about the characteristic function of the leading term. We say that a random vector R_n is $o_D(n^{-\alpha}, n^{-\beta})$ if for any finite constant c,

$$\Pr\left[\|R_n\| > \frac{c}{n^{\alpha}(\log n)^r}\right] = o(n^{-\beta}) \quad \text{for some } r > 0,$$

and use $o_D(n^{-\alpha})$ as shorthand for $o_D(n^{-\alpha}, n^{-\alpha})$. Then, if $T_n = T_n^* + R_n$ with $R_n = o_D(n^{-\alpha})$, and T_n^* having a bounded density uniformly in n, we have

$$\sup_{A \in \mathcal{B}_p} |\Pr[T_n \in A] - \Pr[T_n^* \in A]| = o(n^{-\alpha}),$$

where $\mathcal{B}_p$ consists of all Borel sets in $\mathbb{R}^p$ for which $\mu((\partial A)^{\epsilon}) = O(\epsilon)$ as $\epsilon \to 0$, and we say that T_n is distributionally equivalent to T_n^* to order $n^{-\alpha}$, see Sargan and Mikhail (1971). We also define the asymptotic cumulants of T_n as equal to the cumulants of the random variable T_n^* when the latter has finite moments of the required order.

ASSUMPTION A6. *Suppose that there exists a continuous nonincreasing function* $\chi(t)$ *with* $0 < \chi(t) \leq 1$ *and constant* k *such that* $|\zeta(t)| \leq 1 - \chi(t)$ *for all* $\|t\| > k$, *where* $\zeta(t) = E\left[\exp\left\{\mathbf{i}t' J^{-1} D(X_i)\Omega(X_i)^{-1}\rho(Z_i, \theta_0)\right\}\right]$ *with* $\mathbf{i} = \sqrt{-1}$.

ASSUMPTION A7. *There exists a p-vector of functions* $\psi(\cdot)$, *such that*

$$\widetilde{\theta} - \theta_0 = \frac{1}{n}\sum_{i=1}^{n} \psi(Z_i) + o_D(n^{-2\epsilon}), \tag{8}$$

where $E\left[\psi(Z_i)\right] = 0$ *and* $E\left[\|\psi(Z_i)\|^v\right] < \infty$, $v = 1, 2, \ldots,$ *while* $\epsilon = q/(2q + d)$.

We make all our calculations in the conditional distribution given $\mathcal{X}^n = \{X_1, \ldots, X_n\}$, so that assumptions A6 and A7 are assumed to hold in this distribution with probability one. Assumption A6 is fairly standard in the Edgeworth literature and just rules out certain sorts of discreteness. For a typical parametric estimator, the error in (8) would be $O_p(n^{-1})$, and, when a standard $O(n^{-1/2})$ Edgeworth expansion exists for $\sqrt{n}(\widetilde{\theta} - \theta_0)$, assumption A7 is satisfied for any $\epsilon < 1/2$.

3 Asymptotic Expansions for the Estimator

We shall present our expansions relative to expansions for an infeasible parametric estimator. Let $\overline{\theta}$ be any value that approximately solves $\widetilde{s}(\overline{\theta}) = 0$, where $\widetilde{s}(\theta) = n^{-1}\sum_{i=1}^{n} D(X_i; \widetilde{\theta})\widetilde{\Omega}(X_i; \widetilde{\theta})^{-1}\rho(Z_i, \theta)$; this procedure is infeasible in our case, because it would require a specification for the parametric family $\{D(\cdot;\theta), \Omega(\cdot;\theta); \theta \in \Theta\}$. This estimator is asymptotically normal and is first order equivalent to $\widehat{\theta}$, i.e., $\sqrt{n}(\overline{\theta} - \theta_0) = \sqrt{n}(\widehat{\theta} - \theta_0) + o_p(1)$. Furthermore, $\overline{\theta}$ possesses an Edgeworth expansion to its distribution function under our conditions. Let $\widetilde{F}_{n1}(x)$ be the signed measure with Lebesgue density

$$\widetilde{f}_{n,1}(x) = \phi_{0,I}(x)\left[1 + \frac{\left(H^T\overline{\kappa}\right)^{[3]}(x)}{6\sqrt{n}}\right], \tag{9}$$

where $\left(H^T\overline{\kappa}_n\right)^{[3]}(x)$ is the (third) cumulant weighted sum of a multivariate Hermite polynomial, defined in Barndorff-Nielsen and Cox (1979, p283), which depends on the third (asymptotic) cumulant array $\overline{\kappa}_{n3}$ of $\sqrt{n}(\overline{\theta} - \theta_0)$. Then

$$\sup_{A \in \mathcal{B}_p}\left|\Pr[\sqrt{n}(\overline{\theta} - \theta_0) \in A] - \widetilde{F}_{n,1}(J^{1/2}(A - \mu_{n0}))\right| = o(n^{-2\epsilon}), \tag{10}$$

where the set $J^{1/2}(A - \mu_{n0}) = \{z : z = J^{1/2}(x - \mu_{n0})$ some $x \in A\}$ in which $\mu_{n0} = O(n^{-1/2})$ is the asymptotic mean of the vector $\sqrt{n}(\overline{\theta} - \theta_0)$. See Hall and Horowitz (1996) and Rilestone, Srivastava,

and Ullah (1996). The result (10) follows by standard arguments for nonlinear parametric estimators. First, one approximates the standardized estimator by a polynomial function of a vector of single sums of independent mean zero random variables, and then one uses Edgeworth results for such sequences, see Bhattacharya and Ghosh (1978) for a nice exposition. Likewise, our semiparametric estimator can be approximated by a polynomial function of a vector of [second order] weighted U-statistics. We next present our main result for the estimator $\widehat{\theta}$, which is based on a proof that such random variables have a valid Edgeworth expansion.

Let $B_\Omega(\cdot)$ be the limiting conditional bias of the nonparametric estimator $\widetilde{\Omega}(X_i;\theta_0)$ based on the true parameter, and let $B^\dagger(X) = B(X) - D(X)\Omega(X)^{-1}B_\Omega(X)$. Then define

$$\mu_1 = J^{-1}E\left[\eta_i^\dagger \Omega(X)^{-1}D(X)'J^{-1}D(X)\Omega(X)^{-1}\rho_i\right] \quad ; \quad \Sigma = J^{-1}(\nu_1(\tau)+\nu_2(\tau))J^{-1},$$

where

$$\nu_1(\tau) = \mathfrak{L} - \mathfrak{M}J^{-1}\mathfrak{M}' \quad ; \quad \nu_2(\tau) = \|K\|_2^2 E\left[\frac{S_1(X)+S_2(X,X)}{\tau(X)f^2(X)}\right],$$

where $\mathfrak{L} = E[B^\dagger(X)\Omega(X)^{-1}B^\dagger(X)']$ and $\mathfrak{M} = E[D(X)\Omega(X)^{-1}B^\dagger(X)']$.

THEOREM 1. *Suppose that the regularity conditions A1-A7 hold. Then, there is a sequence of random variables $\widehat{\theta}$ and a constant $k>0$, such that conditional on $\mathcal{X}^n$ with probability one*

$$\Pr\left(\sqrt{n}||\widehat{\theta}-\theta_0|| < k(\log n)^{1/2}, \quad \widehat{\theta} \text{ solves (2)}\right) = 1 - o(n^{-2\epsilon}). \tag{11}$$

Furthermore, there exists sequences of vectors $\{\mu_n\}$ and bounded nonsingular covariance matrices $\{\Psi_n\}$ which are measurable functions of $\mathcal{X}^n$ such that conditional on $\mathcal{X}^n$ with probability one

$$\sup_{A\in\mathcal{B}_p}\left|\Pr[\sqrt{n}(\widehat{\theta}-\theta_0)\in A] - \widetilde{F}_{n,1}(\Psi_n^{-1/2}(A-\mu_n))\right| = o(n^{-2\epsilon}), \tag{12}$$

where $\mu_n = \mu_{n0}+\mu_{n1}$ and $\Psi_n = J^{-1}+\Sigma_n$ such that $n^{1/2}\mu_{n1} = \mu_1 + o_p(n^{1/2-2\epsilon})$ and $n^{2\epsilon}\Sigma_n = \Sigma + o_p(1)$.

The bandwidth and kernel affect only the variance of $\widehat{\theta}$ to the order $n^{-2\epsilon}$; the bias, skewness, and higher cumulants do not depend on bandwidth or kernel to this order. This appears to be a phenomenon common to many adaptive estimators, see for example Linton (1996a) and Xiao and Phillips (1996). Even though neither the bandwidth nor the kernel enter the order $n^{-1/2}$ bias (or skewness) terms, the quantity μ_{n1} exists by virtue of the necessity of estimating the instruments nonparametrically. The signs of the mean corrections μ_{n0} and μ_{n1} are model specific, so that no general conclusions can be drawn about them, while both terms ν_1 and ν_2 are positive, so that

the variance of $\widehat{\theta}$ is unambiguously greater than the limiting value J^{-1} and indeed greater than the variance of $\widetilde{\theta}$.

We conclude this section with an alternative representation of the second-order effect, which follows by an application of deJong's (1987) central limit theorem to the degenerate weighted U-statistics that make up the second order terms.

COROLLARY 1. *As $n \to \infty$, conditional on $\mathcal{X}^n$ with probability one*

$$n^{1/2+\epsilon}(\widehat{\theta} - \overline{\theta}) \Rightarrow N(0, \Sigma).$$

Corollary 1 can serve as the basis for a Hausman test of the null hypothesis that the parametric specification of $\{D(\cdot;\theta), \theta \in \Theta\}$ used in defining $\overline{\theta}$ is correct. Let

$$\mathcal{H} = n^{1+2\epsilon}(\widehat{\theta} - \overline{\theta})^T \widehat{\Sigma}^+ (\widehat{\theta} - \overline{\theta}),$$

where $\widehat{\Sigma}$ is any consistent estimate of Σ and the superscript $+$ denotes generalized inverse. Then, under the null hypothesis $\mathcal{H} \Rightarrow \chi^2(p_0)$, where p_0 is the rank of Σ, while under the alternative, $\mathcal{H} \to \infty$, because then $n^{1/2}(\widehat{\theta} - \overline{\theta}) = O_p(1)$.

4 Testing

We extend the above theory for $\widehat{\theta}$ to the case of semiparametric Wald tests under the null hypothesis. These statistics are based on estimates of the asymptotic covariance matrix, which we treat first.

4.1 Standard Errors

There are many possible estimates of J depending on how and where one substitutes in estimates of the unknown quantities. Let

$$\widetilde{J} = \frac{1}{n} \sum_{i=1}^{n} \widetilde{D}^*(X_i;\widetilde{\theta}) \widetilde{\Omega}(X_i)^{-1} \widetilde{D}^*(X_i;\widetilde{\theta})',$$

where $\widetilde{D}^*(X_i;\widetilde{\theta})$ is like $\widetilde{D}^*(X_i;\widetilde{\theta})$ but constructed with a second bandwidth sequence h^*_{ni}. Note that $\widetilde{J}$ is symmetric and positive semi-definite. Unfortunately, it suffers from a 'degrees of freedom' bias.[3]

[3]This is rather like the maximum likelihood estimator of the variance in a linear regression model.

Specifically, the expansion of $\widetilde{J}$ has a term of order $n^{-1}\underline{h}^{*-d}$ [we suppose that h^*_{ni} is bounded from below and above by sequences $\underline{h}^*(n)$ and $\overline{h}^*(n)$ respectively], which is given by

$$B_J = \frac{1}{n}\sum_{i=1}^{n}\sum_{j\neq i} w_{ij}^2 \Psi_{ij}, \text{ where}$$

$$\Psi_{ij} = E[\eta_j \Omega(X_i)^{-1}\eta_j' - \eta_j\Omega(X_i)^{-1}\zeta_j\Omega(X_i)^{-1}D(X_i)' - D(X_i)\Omega(X_i)^{-1}\zeta_j\Omega(X_i)^{-1}\eta_j'|X_i, X_j].$$

We propose to make a multiplicative bias correction to $\widetilde{J}$ that eliminates this term; specifically, we take

$$\widehat{J} = \widetilde{J}^{1/2}\exp(-\widetilde{J}^{-1/2}\widetilde{B}_J\widetilde{J}^{-1/2})\widetilde{J}^{1/2},$$

where exp is the matrix exponential of a real symmetric matrix, and

$$\widetilde{B}_J = \frac{1}{n}\sum_{i=1}^{n}\sum_{j\neq i} w_{ij}^2[\widetilde{\eta}_j\widetilde{\Omega}(X_i)^{-1}\widetilde{\eta}_j' - \widetilde{\eta}_j\widetilde{\Omega}(X_i)^{-1}\widetilde{\zeta}_j\widetilde{\Omega}(X_i)^{-1}\widetilde{D}(X_i)' - \widetilde{D}(X_i)\widetilde{\Omega}(X_i)^{-1}\widetilde{\zeta}_j\widetilde{\Omega}(X_i)^{-1}\widetilde{\eta}_j'],$$

where $\widetilde{\eta}_j$ and $\widetilde{\zeta}_j$ are the corresponding residuals. The resulting covariance matrix estimator $\widehat{J}$ is symmetric and positive semi-definite, and has smaller order (degrees of freedom) bias than $\widetilde{J}$; this can be seen from a Taylor series expansion argument.

Note that there is a different trade-off between bias and variance terms in $\widehat{J}$ than in $\widehat{\theta}$, which is what motivates the different bandwidth h^*_{ni}.

4.2 Wald Statistic

Suppose we wish to test the following nonlinear hypothesis concerning the parameters

$$\mathrm{H}_0\text{: } g(\theta_0) = 0 \quad ; \quad \mathrm{H}_A\text{: } g(\theta_0) \neq 0,$$

where $g(\cdot)$ is a $p_1 \times 1$ vector of continuously differentiable functions with $p_1 \leq p$. The properties of the parametric Wald statistic

$$\overline{W} = n\overline{g}'\{\overline{GJ}^{-1}\overline{G}'\}^{-1}\overline{g},$$

where $\overline{g} = g(\overline{\theta})$ and $\overline{G} = G(\overline{\theta})$ with $G(\theta) = \partial g(\theta)/\partial\theta'$ and $\overline{J} = n^{-1}\sum_{i=1}^{n} D(X_i;\widetilde{\theta})\widetilde{\Omega}(X_i)^{-1}D(X_i;\widetilde{\theta})'$, are well known. Specifically, under regularity conditions

$$\Pr\left[\overline{W} > \chi^2_{p_1}(\alpha)\,|\mathrm{H}_0\right] = \alpha + o(1) \quad ; \quad \Pr\left[\overline{W} > \chi^2_{p_1}(\alpha)\,|\mathrm{H}_A\right] = 1 + o(1), \tag{13}$$

where $\chi^2_p(\alpha)$ denotes the α^{th} critical value of the chi-squared(p) random variable χ^2_p. In fact, $\kappa_j(\overline{W}) = \kappa_j(\chi^2_{p_1}) + O(n^{-1})$, $j = 1, \ldots,$ and, furthermore, the errors in (13) are actually $O(n^{-1})$. Phillips and Park (1988) compute explicitly the order n^{-1} correction terms in some special cases.

We consider the semiparametric Wald statistic

$$\widehat{W} = n\widehat{g}'\{\widehat{G}\widehat{J}^{-1}\widehat{G}'\}^{-1}\widehat{g}, \tag{14}$$

where $\widehat{g} = g(\widehat{\theta})$ and $\widehat{G} = G(\widehat{\theta})$. Under our regularity conditions, $\widehat{W}$ and $\overline{W}$ are first-order equivalent under the null hypothesis, but not second-order equivalent. Our expansions reveal that if the same bandwidth magnitude $h_{ni} = O(n^{-1/(2q+d)})$ is used in estimating θ and J, then the bias from estimating J contributes a large second order effect to $\widehat{W}$. To avoid this, we suppose that h_{ni} satisfying A5 is used to construct $\widehat{\theta}$, while a second bandwidth h^*_{ni} is used in $\widehat{J}$, i.e., $\widehat{J}(h^*_{ni})$, where h^*_{ni} is smaller in order than h_{ni}. In this case, $\widehat{W}(h_{ni}, h^*_{ni})$ has second order effect which is the same magnitude as in estimation, and is indeed smaller than can be achieved when a single bandwidth is used throughout.

THEOREM 2. *Suppose that the regularity conditions A1-A7 hold, that the function g is four times continuously differentiable in a neighborhood of θ_0, and that the matrix $Q = G_0'\{G_0 J^{-1} G_0'\}^{-1} G_0$, where $G_0 = G(\theta_0)$, is of full rank. Suppose that $n^2 \overline{h}_n^{*2q+d} \to 0$ and $n^2 \underline{h}^{*2d/(q+d)} \to \infty$. Then, conditional on $\mathcal{X}^n$ with probability one, we have*

$$\Pr\left[\widehat{W} \leq x \,|\mathrm{H}_0\right] = F_{p_1,\infty}\left[x(1 - \mathrm{tr}(\Xi_n))\right] + o(n^{-2\epsilon}), \tag{15}$$

where $\Xi_n = 2Q\Sigma_n/p_1$ with Σ_n as defined below Theorem 1.

REMARKS.

1. The null rejection frequency of the test based on the asymptotic critical values $\chi^2_{p_1}(\alpha)$ is $\alpha + O(n^{-2\epsilon})$, where the $O(n^{-2\epsilon})$ term depends on the bandwidth constant through the matrix Σ_n. Since Σ_n is positive semi-definite, the test based on the asymptotic critical values will tend to over-reject. The precise magnitude of the second order effect depends also on the restriction through the matrix Q.

2. When a single bandwidth is used throughout, i.e., $h^*_{ni} = h_{ni}$, then the order of magnitude of the correction term in (15) is larger; specifically it is of order $n^{-q/(q+d)}$ when $h_{ni} = O(n^{-1/(q+d)})$. Furthermore, the direction of the effect could take either sign.

5 Second-Order Efficiency and Bandwidth Selection

5.1 Optimal Estimation

We now suppose that the bandwidth is of the form $h_n(x) = \tau_\lambda(x) n^{-1/(2q+d)}$ for a family of functions $\tau_\lambda(\cdot)$, $\lambda \in \Lambda \subseteq \mathbb{R}^l$. For example, the family $\tau_\lambda(x) = \lambda_1 f^{-\lambda_2}(x)$ contains fixed window estimation [$\lambda_2 = 0$] and 'ideal' nearest neighbor estimation [$\lambda_2 = 1$] as well as a range of intermediate schemes. Jennen-Steinmetz and Gasser (1988) discuss this bandwidth scheme and its implications at some length. Our objective is to define a general optimality criterion for deciding on a best choice of the parameter λ for estimation and testing purposes. Define by $\mathcal{C}$ the class of estimators $\widehat{\theta}$ defined by (2) which are based on the smoothing procedure (5) with bandwidth sequence as above. Each member of $\mathcal{C}$ is indexed by a value of λ; we shall seek the 'best' value, which we denote by λ_{opt}.[4]

Recall that the (asymptotic) squared bias of the estimator is of order n^{-1} and so the (asymptotic) mean squared error is the same as the asymptotic variance to order $n^{-2\epsilon}$. We shall work with a scalar valued risk function R that depends only on the asymptotic variance matrix Ψ_n of the estimator, i.e., $R(\widehat{\theta}, \theta_0) = \varrho(\Psi_n)$, where ϱ is some smooth scalar valued function. For example, ϱ could be any of the widely used criteria for multivariate optimality such as the determinant or trace of the matrix or a particular quadratic form $c'\Psi_n c$. The covariance matrix Ψ_n depends on λ only through the higher order terms, i.e.,

$$\Psi_n(\lambda) = J^{-1} + \Sigma_n(\lambda) := J^{-1} + \Sigma_{n1}(\lambda) + \Sigma_{n2}(\lambda),$$

where $\Sigma_{n1}(\lambda) = n^{-2\epsilon} J^{-1} \nu_1(\tau_\lambda) J^{-1}$ and $\Sigma_{n2}(\lambda) = n^{-2\epsilon} J^{-1} \nu_2(\tau_\lambda) J^{-1}$. Because the distribution of $\sqrt{n}(\widehat{\theta} - \theta_0)$ only depends on the bandwidth through Ψ_n [to this order] there is no loss of generality in choosing this criterion if the underlying loss function is symmetric about zero and bowl-shaped. See Rothenberg (1984, pp902-909) for some discussion about higher order efficiency. A linear approximation to $\varrho(\Psi_n)$ ignoring terms that don't depend on λ is

[4] More broadly, though, one might want to compare methods in the broader class where $\tau(\cdot)$ is considered an unknown function. It may be possible to find an optimal $\tau(\cdot)$ through calculus of variation techniques, but it will depend on the data distribution in a complicated fashion. It is easy to see that no particular $\tau(\cdot)$ method can dominate any other according to ϱ uniformly over the class of joint distributions $\mathcal{D}$ of the data Z_i, $i = 1, \ldots, n$.

One way of obtaining an unambiguous ranking might be through the minimax criterion, as in Fan (1993). However, it is to be expected that no method can be found to achieve this bound. In any case, this way of comparing estimators is of questionable value, since it makes nature seem unduly hostile. Herman Chernoff [Bather (1997, p. 339)] gives the following example: "If you have a choice between committing suicide or not taking action, in which case you might lead a normal life or you might die a horrible death, the minimax principle tells you to avoid any possibility [however remote] of a horrible death by committing suicide."

$$\sigma_n(\lambda) = \varrho'\omega_1(\tau_\lambda) + \varrho'\omega_2(\tau_\lambda), \tag{16}$$

where $\varrho = \{\partial\varrho\,/\mathrm{vech}(\Psi')\}_{\Psi=J^{-1}}$, $\omega_1(\tau_\lambda) = \mathrm{vech}(J^{-1}\nu_1(\tau_\lambda)J^{-1})$, and $\omega_2(\tau_\lambda) = \mathrm{vech}(J^{-1}\nu_2(\tau_\lambda)J^{-1})$. Minimizing $\varrho\{\Psi_n(\lambda)\}$ is equivalent to minimizing $\sigma_n(\lambda)$. We shall suppose that $0 < \underline{\tau} = \inf_{\lambda,x}\tau_\lambda(x) \leq \sup_{\lambda,x}\tau_\lambda(x) = \overline{\tau} < \infty$ and that there is a unique λ, denoted λ_{opt}, that minimizes $\sigma(\lambda)$ over the set Λ. Finally, let

$$h_{opt}(x) = \tau_{\lambda_{opt}}(x)n^{-1/(2q+d)}. \tag{17}$$

The resulting estimator is the best within the class $\mathcal{C}$.[5] In the special case that $\tau_\lambda(x) = \lambda\tau_0(x)$ for some fixed function $\tau_0(x)$, λ_{opt} has an explicit solution. i.e., $\lambda_{opt} = \{(d\varrho'\omega_2)/(2q\varrho'\omega_1)\}^{1/(2q+d)}$. More generally, one must solve a nonlinear optimization problem to find λ_{opt}.

5.2 Feasible Bandwidth Selection

Let $\widehat{\sigma}_n(\lambda)$ be some estimate of $\sigma_n(\lambda)$, let $\widehat{\lambda}_{opt}$ minimize $\widehat{\sigma}_n(\lambda)$ with respect to $\lambda \in \Lambda$, and let

$$\widehat{h}_{opt}(x) = \tau_{\widehat{\lambda}_{opt}}(x)n^{-1/(2q+d)}.$$

This data-based bandwidth selection method mimics (17). Provided $\widehat{\sigma}_n(\lambda)$ is a good approximation to $\sigma_n(\lambda)$ we can expect that $\widehat{h}_{opt}$ will be a good approximation to h_{opt}. Specifically, if

$$\sup_{\lambda\in\Lambda}|\widehat{\sigma}_n(\lambda) - n^{-2\epsilon}\sigma(\lambda)| = o_p(n^{-2\epsilon}),$$

then

$$h_{opt}^{-1}(\widehat{h}_{opt} - h_{opt})(x) = o_p(1)$$

uniformly in x. Under additional conditions we can establish the second order efficiency of the estimator $\widehat{\theta}$ based on data-based bandwidth selection. Below we give such a result under high level conditions.

THEOREM 3. *Suppose that for some φ_1 and φ_2 with $0 < \varphi_1, \varphi_2$, where $\varphi_1 + \varphi_2 > 2\epsilon + 1/2$, then*

$$\widehat{\lambda}_{opt} - \lambda_{opt} = o_D(n^{-\varphi_1}, n^{-2\epsilon}) \tag{18}$$

$$\sup_{\|\lambda-\lambda_{opt}\|\leq\delta n^{-\varphi_1}} \left\|\frac{\partial\widehat{\theta}}{\partial\lambda}(\lambda)\right\| = o_D(n^{-\varphi_2}, n^{-2\epsilon}). \tag{19}$$

[5] By contrast, the cross-validation bandwidth selection method proposed in Newey (1990) achieves the right bandwidth rate, although the constants would generally be wrong and would lead to suboptimal risk according to our criterion.

Then, the distribution of $\sqrt{n}(\widehat{\theta}(\widehat{\lambda}_{opt}) - \theta_0)$ *is the same as the distribution of* $\sqrt{n}(\widehat{\theta}(\lambda_{opt}) - \theta_0)$ *to order* $n^{-2\epsilon}$, *i.e.,* $\widehat{\theta}(\widehat{\lambda}_{opt})$ *is second order efficient.*

The condition (18) requires that $\widehat{\lambda}_{opt}$ be a little more than n^{φ_1}-consistent. It can be shown [for $\varphi_1 < 1/2$] by using the same arguments under additional smoothness conditions on the data distribution. The condition (19) requires that the effect of λ on $\widehat{\theta}(\lambda)$ is small in a certain sense. It is like a stochastic equicontinuity condition. It follows from the same sort of arguments used in establishing Theorem 1. See Linton (1995) for similar arguments.

We now discuss estimation of $\sigma_n(\lambda)$. It may appear that since $\sigma_n(\lambda)$ depends on the unknown derivatives of D etc, any method for estimating $\sigma_n(\lambda)$ must involve additional nonparametric estimation of these unknown quantities, which presupposes the selection of a preliminary bandwidth. We argue that this is not really the case; one can make any of these methods depend only on a single unknown quantity λ by the well respected method of profiling or concentration, see Jones, Marron, and Sheather (1992) for a review of such methods in kernel density estimation. That is, we use nonparametric estimators of $\sigma_n(\lambda)$ that depend on a smoothing parameter which is also controlled by the same λ. Provided the magnitude of these bandwidths is chosen correctly, the resulting estimator will have the required properties. Specifically, let

$$\widehat{\Sigma_{n1}(\lambda)} = \widehat{J}^{-1}(\widehat{\mathfrak{L}} - \widehat{\mathfrak{M}}\widehat{J}^{-1}\widehat{\mathfrak{M}}')\widehat{J}^{-1} \tag{20}$$

$$\widehat{\Sigma_{n2}(\lambda)} = \widehat{J}^{-1}\{\frac{1}{n}\sum\sum_{j\neq i} w_{ij}^2\widetilde{\eta}_j^{\dagger}\widetilde{\Omega}(X_i)^{-1}\widetilde{\eta}_j^{\dagger\prime} + \sum\sum_{j\neq i} w_{ij}w_{ji}\widetilde{\eta}_j^{\dagger}\widetilde{\Omega}(X_i)^{-1}\widetilde{\rho}_i\widetilde{\rho}_j'\widetilde{\Omega}(X_i)^{-1}\widetilde{\eta}_i^{\dagger\prime}\}\widehat{J}^{-1}, \tag{21}$$

where $\widehat{\mathfrak{L}} = n^{-1}\sum_{i=1}^n \widetilde{B}_{ni}^{\dagger}\widetilde{\Omega}(X_i)^{-1}\widetilde{B}_{ni}^{\dagger\prime}$ and $\widehat{\mathfrak{M}} = n^{-1}\sum_{i=1}^n \widetilde{D}(X_i;\widetilde{\theta})\widetilde{\Omega}(X_i)^{-1}\widetilde{B}_{ni}^{\dagger\prime}$, where $\widetilde{B}_{ni}^{\dagger}$ is some nonparametric estimator of $B_{ni}^{\dagger}$, while $\widetilde{\eta}_j^{\dagger}$ are residuals corresponding to $\eta_j^{\dagger}$. Finally, let $\widehat{\lambda}$ minimize

$$\widehat{\sigma}_n(\lambda) = \widehat{\varrho}'\widehat{\omega}_{n1}(\lambda) + \widehat{\varrho}'\widehat{\omega}_{n2}(\lambda), \tag{22}$$

where $\widehat{\varrho} = \{\partial\varrho/\mathrm{vech}(\Psi)\}_{\Psi=\widehat{J}^{-1}}$, while $\widehat{\omega}_{n1}(\lambda) = \mathrm{vech}(\widehat{\Sigma_{n1}(\lambda)})$ and $\widehat{\omega}_{n2}(\lambda) = \mathrm{vech}(\widehat{\Sigma_{n2}(\lambda)})$, and let

$$\widehat{h}_n(x) = \tau_{\widehat{\lambda}}(x)n^{-1/(2q+d)}.$$

The main issue arises with how to construct $\widetilde{B}_{ni}^{\dagger}$, since all the other quantities in (20) and (21) can be computed from what we know already. We just outline the procedure for $\widetilde{B}_{ni}$, but the same applies to $\widetilde{B}_{\Omega ni}$ and hence $\widetilde{B}_{ni}^{\dagger}$. Recall that $B_{ni} = \sum_{j\neq i} w_{ij}\{D(X_j) - D(X_i)\}$; this suggests a general method for estimating the bias. Specifically, we take

$$\widetilde{B}_{ni} = \sum_{j \neq i} w_{ij} \{ \widehat{D}(X_j) - \widehat{D}(X_i) \}, \tag{23}$$

where the weights $\{w_{ij}\}$ are precisely those used in (3), while $\widehat{D}(X_j)$ is some estimate of $D(X_j)$. For example, let

$$\widehat{D}(X_i) = \sum_{j \neq i} w^*_{ij} \frac{\partial \rho(Z_j, \widetilde{\theta})}{\partial \theta'},$$

where the weights $\{w^*_{ij}\}$ come from a local polynomial regression of order $q+s$ for some $s > 0$, and have bandwidth sequence

$$h^B_n(x) = \tau_\lambda(x) n^{-1/(2q+2s+d)}.$$

It is important to notice here that although the rate of $h^B_n(x)$ is different from the rate of $h_n(x)$, these bandwidths are both determined by the same unknown parameter λ. Provided D and f are smooth enough, i.e., satisfy A3 with $q+s$ replacing q, the estimation error in $\widetilde{B}_{ni}$ will be smaller than B_{ni} itself. An alternative method called the 'rule of thumb' approach is based on a parametric specification of D. This method was proposed in Silverman (1986) and further exploited in Andrews (1991) in other contexts. In our case, one takes an auxiliary parametric specification $\{D(\cdot;\vartheta), \vartheta \in \Upsilon\}$ for $D(\cdot)$, and computes estimates $\widetilde{\vartheta}$ of ϑ by some parametric estimation technique like maximum likelihood. We then take $D(\cdot;\widetilde{\vartheta})$ in place of $\widehat{D}(\cdot)$ in (23) and hence (20).[6] We investigate both the nonparametric plug-in and the rule-of-thumb methods in the Monte Carlo experiments below. The 'variance' terms in (21) like $\sum_{j \neq i} \sum w^2_{ij} \widetilde{\eta}^\dagger_j \widetilde{\Omega}(X_i)^{-1} \widetilde{\eta}^{\dagger\prime}_j / n$ well approximate their estimands as can be shown by standard arguments for U-statistics, see Fan and Li (1996).

Finally, we show how optimal testing can be put in the same framework. We shall suppose now that $h_n(x) = \tau_\lambda(x) n^{-1/(2q+d)}$ is used in $\widehat{\theta}$, while $h^*_n(x) = \tau_\lambda(x) n^{-1/(2q+d)} n^{-\delta}$ for some $\delta > 0$, is used in $\widehat{J}$ as required by the theorem. Both the rates $n^{-1/(2q+d)}$ and $n^{-1/(2q+d)} n^{-\delta}$ are given and the only unknown quantity is λ, whose value will be determined in the sequel. Suppose that we define an optimal bandwidth for $\widehat{W}$ as one that produces the smallest discrepancy between the actual null rejection frequency and the nominal level of the test, that is, we choose the bandwidth constant to solve the following problem

$$\min_\lambda \left| \Pr \left[\widehat{W} > \chi^2_{p_1}(\alpha) \,|\mathrm{H}_0 \right] - \alpha \right|. \tag{24}$$

This criterion is not the only one that one might be interested in, but it is quite important in itself. Furthermore, it does at least prove to be tractable. This concentration on the null hypothesis is

[6] Under suitable conditions, this method provides second-order optimality for the resulting $\widehat{\theta}$ when $D(\cdot)$ is as specified, and the right magnitude for h more generally.

very similar to that adopted in the bootstrap literature, see Hall (1992, p222). By Theorem 2 and a Taylor expansion, we have

$$\begin{aligned}\Pr\left[\widehat{W} > \chi^2_{p_1}(\alpha)\,|\mathrm{H}_0\right] - \alpha &= 1 - F_{p_1,\infty}\left[\chi^2_{p_1}(\alpha)(1-\chi^2_{p_1}(\alpha)\mathrm{tr}(\Xi_n))\right] - \alpha + o(n^{-2\epsilon}) \\ &= f_{p_1,\infty}\left[\chi^2_{p_1}(\alpha)\right]\chi^2_{p_1}(\alpha)\mathrm{tr}(\Xi_n) + o(n^{-2\epsilon}).\end{aligned}$$

Therefore, since $\chi^2_{p_1}(\alpha) > 0$ and $f_{p_1,\infty}(\chi^2_{p_1}(\alpha)) < 0$ always, the optimal bandwidth according to (24) equivalently minimizes $\mathrm{tr}(\Xi_n(\lambda))$. It is the same magnitude as the optimal bandwidth in the estimation problem; it depends on the restrictions through G_0, but does not depend on α. Note that we can write $\mathrm{tr}(\Xi_n(\lambda))$ as $\varrho(\Psi_n(\lambda))$ for some function ϱ, which puts the estimation and testing problems within a common framework. In the testing case, a feasible bandwidth selection method can be based on replacing $\Xi_n(\lambda)$ by an estimate $\widehat{\Xi}_n(\lambda) = 2\widehat{G}\widehat{\Sigma}_n(\lambda)\widehat{G}'\{\widehat{G}\widehat{J}^{-1}\widehat{G}'\}^{-1}/p_1$ and proceeding as above.

6 Monte Carlo Experiment

We evaluated our second-order approximations and the bandwidth selection procedure on the sample selection model given below. We report here only results about testing, see Linton (1997) for results on estimation.

$$Y_i = c + \beta_{00}s_i + \sum_{j=1}^{4}\beta_{j0}X_{ji} + \varepsilon_i \quad ; \quad s_i = \mathbf{1}\left(\alpha_{10} + \alpha_{20}X_{0i} + \eta_i > 0\right),$$

$$\begin{bmatrix}\varepsilon_i \\ \eta_i\end{bmatrix} \sim N\left(\begin{bmatrix}0 \\ 0\end{bmatrix}, \begin{bmatrix}1 & 0.7 \\ 0.7 & 1\end{bmatrix}\right),$$

where $(X_{0i}, X_{1i}, \ldots, X_{4i})$ are mutually independent with marginal distributions standard normal truncated [at +/-3]. We take $\alpha_{10} = \alpha_{20} = 1$ throughout. We computed the Wald statistics for testing the two hypotheses: (a) $\mathbf{H}_0$: $\beta_{00} = 0$ [in which case we set $\beta_{00} = 0$ and $\beta_{j0} = 1$, $j = 1, \ldots, 4$.]; (b) $\mathbf{H}_0$: $\beta_{j0} = 0$, $j = 1, \ldots, 4$ [in which case we set $\beta_{00} = 1$ and $\beta_{j0} = 0$, $j = 1, \ldots, 4$.]. We shall consider the constant design [in which $c = 1$], and the no-constant design [in which $c = 0$ and is not estimated]. In total this makes four designs.

We suppose that it is known that selection is determined only by X_0, in which case the optimal instrument for s is $\pi(x_0) = \Pr[s = 1|X_0 = x_0]$. We estimate this function by the Nadaraya-Watson

estimator and the local linear method using the following bi-quadratic kernel function: $K(u) = 0.9375(1-u^2)^2\mathbf{1}(|u| \leq 1)$ and a constant bandwidth $h_n = \lambda n^{-1/5}$ when used in computing $\widehat{\theta}$ and $h_n^* = \lambda^* n^{-1/2}$ when used in computing $\widehat{J}_1$ in $\widehat{W}$. We computed the test statistic at a grid of fixed bandwidths with λ in the range $[\lambda_{\min}, \lambda_{\max}]$, taking $\lambda^* = 0.9 * \lambda_{\min}$ throughout. We also computed the test statistics using our data-based bandwidth selection method to determine λ. Specifically, we estimated the bias by (23) using two different methods to compute $\widehat{D}(\cdot)$, which in our case is $\widehat{\pi}(\cdot)$: nonparametric plug-in and rule-of-thumb plug-in. In the former case, we took $\widetilde{\pi}(\cdot)$ as our estimate of $\pi(x)$ in (23). In the latter case, we take a parametric specification $\{\pi_p(\cdot;\vartheta),\, \vartheta \in \Upsilon\}$ for $\pi(\cdot)$, and compute estimates $\widetilde{\vartheta}$ of ϑ by maximum likelihood. We then take $\pi_p(\cdot;\widetilde{\vartheta})$ in place of $\widehat{D}(\cdot)$ in (23) and hence (20).[7] We take π_p to be either a quadratic function or a logit with linear index.

Throughout, we examine the conditional distribution of $\widehat{W}\,|X_1,\ldots,X_n$. The results of $20,000$ replications are given in Figures 1-3 below. First, there is quite substantial over-rejection which decreases with sample size for each bandwidth value. There is some variation in the bandwidth effect across designs. While in Figure 1 there is a substantial effect, in Figure 2, where there is no constant, the rejection frequency is much less sensitive to bandwidth. Figure 3 is somewhat intermediate, but still exhibits less dependence on bandwidth. The main reason for this lack of sensitivity is that in these cases the matrix J^{-1} is approximately diagonal. Since the bandwidth effects only come in from the estimation of $\pi(X_{0i})$, i.e., the matrix $\mathfrak{L}_n$ is zero except for the corresponding diagonal element, our theory does predict that the second order matrix Σ should be zero or approximately zero in this case.

We now turn to Figure 1. The local linear fixed bandwidth results show a poorer performance at small bandwidths than the kernel method, while at larger bandwidth the local linear does better. The automatic bandwidth selection methods generally do pretty well, being in some cases slightly better than the best fixed bandwidth and in some cases slightly worse. There is not much to choose amongst the different bandwidth selection methods, although the nonparametric method appears to do the best and the quadratic rule-of-thumb method does the worst. We also compared the distribution of the Wald statistic with the corresponding chi-squared distribution (not shown) - the distributions are quite close.

[7] Under suitable conditions, this method provides second-order optimality for the resulting $\widehat{\theta}$ when $D(\cdot)$ is as specified, and the right magnitude for h more generally. This method was proposed in Silverman (1986) and further exploited in Andrews (1991) in other contexts.

7 Conclusion

In concluding, we mention some extensions of this paper. Our theory is restricted to the case where the marginal density of the covariates is bounded away from zero on its compact support. When this assumption is violated, a totally new theory is necessary which would combine extreme value theory with the methods we use here. This remains to be addressed in future work. Also, the naive bootstrap, which draws samples $\{Y_i^*, X_i^*\}_{i=1}^n$, although first order correct, will fail to capture the second order effects. Specifically, this bootstrap will fail to capture the bias-related terms, as it does in nonparametric density and regression problems, see Härdle and Marron (1991). In order to provide good approximation at the optimal bandwidth it is necessary to use a more complicated bootstrap algorithm.

A Appendix

The appendices are organized as follows. In section A1 we give some background results on kernel regression estimation. In section A2 we establish an approximation for the first three derivatives of the score function. In appendix B we present the proofs of our theorems. In appendix C we give the proofs of six lemmas given in Appendix A.

Throughout let k denote a positive finite constant which can be different from expression to expression.

A.1 Nonparametric Preliminaries

Let $t(q,d) = \sum_{\ell=0}^{q-1} t_\ell$ denote the total number of parameters in the vector $\boldsymbol{\alpha}(X_i)$, where $t_\ell = \binom{\ell+d-1}{d-1}$ is the total number of distinct partial derivatives of order ℓ. We will assume that these partial derivatives have been arranged in some order, which will be the same throughout the sequel. Define the real symmetric $t \times t$ matrices

$$M_{ni} = \begin{bmatrix} M_{ni,0,0} & M_{ni,0,1} & \cdots & M_{ni,0,q} \\ M_{ni,1,0} & M_{ni,1,1} & & \vdots \\ \vdots & & \ddots & \\ M_{ni,q,0} & M_{ni,q,1} & \ldots & M_{ni,q,q} \end{bmatrix} \quad ; \quad M_i = f(X_i) \begin{bmatrix} M_{0,0} & M_{0,1} & \cdots & M_{0,q} \\ M_{1,0} & M_{1,1} & & \vdots \\ \vdots & & \ddots & \\ M_{q,0} & M_{q,1} & \ldots & M_{q,q} \end{bmatrix},$$

in which each sub-matrix $M_{ni,\ell,k}$ [and $M_{\ell,k}$] has dimensions $t_\ell \times t_k$. The typical element of the sub-matrix $M_{ni,\ell,k}$ is $m_{n,\mathbf{a}+\mathbf{b}}(X_i)$, where $|\mathbf{a}| = \ell$ and $|\mathbf{b}| = k$, in which for each d-tuple $\mathbf{a} = (a_1, \ldots, a_d)$

$$m_{n,\mathbf{a}}(X_i) = \frac{1}{nh_{ni}^d} \sum_{j \neq i} K\left(\frac{X_i - X_j}{h_{ni}}\right) \left(\frac{X_i - X_j}{h_{ni}}\right)^{\mathbf{a}},$$

while the typical element in $M_{j,k}$ is $m_{\mathbf{a}+\mathbf{b}} = \int u^{\mathbf{a}+\mathbf{b}} K(u)du$. Define the local polynomial weights as

$$w_{ij} = \sum_{\{\mathbf{a}:|\mathbf{a}|\leq q\}} w_{ij}^{\mathbf{a}} \quad ; \quad w_{ij}^{\mathbf{a}} = \frac{1}{nh_{ni}^d} e_1' M_{ni}^{-1} e_{\ell(\mathbf{a})} K\left(\frac{X_i - X_j}{h_{ni}}\right) \left(\frac{X_i - X_j}{h_{ni}}\right)^{\mathbf{a}},$$

where for each vector $\mathbf{a}$ with $|\mathbf{a}| \leq q$, $\ell(\mathbf{a})$ is its position in the $t \times 1$ vector $\boldsymbol{\alpha}(X_i)$, and $e_j = (0, \ldots, 1, \ldots 0)$ is the j^{th} elementary vector. Define also the $t \times t$ matrix $\overline{M}_{ni} = E(M_{ni}|X_i)$. Let $B_{\Omega ni} = \sum_{j \neq i} w_{ij}\Omega(X_j) - \Omega(X_i)$, $V_{ni} = \sum_{j \neq i} w_{ij}\eta_j$, and $V_{\Omega ni} = \sum_{j \neq i} w_{ij}\zeta_j$.

We make use of the following lemmas.

LEMMA 1. *For all $\alpha, \beta = 1, \ldots, p$, with probability one,*

$$\max_{1 \leq i \leq n} \left| \left(M_{ni} - \overline{M}_{ni}\right)_{\alpha\beta} \right| = O\left(\frac{(\log n)^{1/2}}{(n\underline{h}^d)^{1/2}}\right) \tag{25}$$

$$\max_{1 \leq i \leq n} \left| \left(\overline{M}_{ni} - M_i\right)_{\alpha\beta} \right| = O(\overline{h}) \tag{26}$$

$$\max_{1 \leq i \leq n} \left| \widetilde{D}_{\alpha\beta}(X_i; \theta_0) - D_{\alpha\beta}(X_i) \right| = O\left(\frac{(\log n)^{1/2}}{(n\underline{h}^d)^{1/2}}\right) + O\left(\overline{h}^q\right). \tag{27}$$

LEMMA 2. *Suppose that the conditions A1-A5 hold except that we only require ν moments with $\nu \geq (1+\epsilon)/\epsilon$. Then, for some finite k,*

$$\Pr\left(\max_{1 \leq i \leq n} |(M_{ni} - M_i)_{\alpha\beta}| > k\frac{\log n}{n^\epsilon}\right) = o(n^{-2\epsilon}) \tag{28}$$

$$\Pr\left(\max_{1 \leq i \leq n} |(B_{ni})_{\alpha\beta}| > k\frac{\log n}{n^\epsilon}\right) = o(n^{-2\epsilon}) \tag{29}$$

$$\Pr\left(\max_{1 \leq i \leq n} |(V_{ni})_{\alpha\beta}| > k\frac{\log n}{n^\epsilon}\right) = o(n^{-2\epsilon}). \tag{30}$$

This says that the corresponding random sequence is $o_D(n^{-(\epsilon-\eta)}, n^{-2\epsilon})$ for any $\eta > 0$.

The results (27), (29) and (30) are also true for $\widetilde{\Omega}(X_i; \theta_0)$, $B_{\Omega ni}$ and $V_{\Omega ni}$.

LEMMA 3. *With probability one, as* $n \to \infty$,

$$\begin{aligned} \max_{1\le i\le n} \#\{j : w_{ij} \neq 0\} &= O(n^{2\epsilon}) && (31)\\ \max_{1\le i,j\le n} |w_{ij}| &= O(n^{-2\epsilon}) && (32)\\ \max_{1\le i\le n} \sum_{j\neq i} |w_{ij}| &= O(1). && (33) \end{aligned}$$

There exists a matrix of random variables $\Psi(Z_i)$ whose elements have mean zero [with probability one conditional on X_i] and finite second moments such that

$$\widetilde{\Omega}(X_i) - \Omega(X_i) = V_{\Omega ni} + B_{\Omega ni} + \frac{1}{n}\sum_{i=1}^{n} \Psi(Z_i) + o_D(n^{-2\epsilon}) \tag{34}$$

by Taylor expansion and Assumptions A2 and A7. Similarly,

$$\widetilde{D}(X_i;\widetilde{\theta}) - D(X_i) = V_{ni} + B_{ni} + \sum_{\gamma=1}^{p} F_\gamma(X_i)\frac{1}{n}\sum_{j=1}^{n} \psi_\gamma(Z_j) + o_D(n^{-2\epsilon}), \tag{35}$$

where $F_\gamma(X_i)$ is the $p \times p$ matrix with typical elements $F_{\gamma;\alpha\beta}(X_i) = E[\partial^2\rho_\alpha(Z_i,\theta_0)/\partial\theta_\beta\partial\theta_\gamma \,|X_i]$, for $\alpha,\beta,\gamma = 1,\ldots,p$.

A.2 Standardized Criterion Derivatives

We present here the three crucial lemmas that are used in the proofs of Theorem 1 and 2

LEMMA 4. *There exist random vectors* S_j, $j = 0,\ldots,4$, *with* $E[\|S_j\|^v] < \infty$ *for* $v = 1,2,\ldots,$ *and* $j = 0,\ldots,4$, *such that*

$$\sqrt{n}\widehat{s}(\theta_0) = \sum_{j=0}^{4} S_j + o_D(n^{-2\epsilon}),$$

where

$$\begin{aligned} S_0 &= \frac{1}{\sqrt{n}}\sum_{i=1}^{n} D(X_i)\Omega(X_i)^{-1}\rho(Z_i,\theta_0) = O_p(1)\\ S_1 &= -\frac{1}{n}\sum_{i=1}^{n} D(X_i)\Omega(X_i)^{-1}\{\frac{1}{\sqrt{n}}\sum_{l=1}^{n}\Psi(Z_l)\}\Omega(X_i)^{-1}\rho(Z_i,\theta_0) = O_p(n^{-1/2})\\ S_2 &= \frac{1}{n}\sum_{j=1}^{n}\sum_{\gamma=1}^{p}\psi_\gamma(Z_j)\frac{1}{\sqrt{n}}\sum_{i=1}^{n} F_\gamma(X_i)\Omega(X_i)^{-1}\rho(Z_i,\theta_0)\} = O_p(n^{-1/2}) \end{aligned}$$

$$
\begin{aligned}
S_3 &= \frac{1}{\sqrt{n}} \sum_{i=1}^{n} B_{ni}^{\dagger} \Omega(X_i)^{-1} \rho(Z_i, \theta_0) = O_p(\overline{h}^q) \\
S_4 &= \frac{1}{\sqrt{n}} \sum_{i=1}^{n} V_{ni}^{\dagger} \Omega(X_i)^{-1} \rho(Z_i, \theta_0) = O_p(n^{-1/2} \underline{h}^{-d/2}),
\end{aligned}
$$

where $B_{ni}^{\dagger} = B_{ni}^{\dagger} - D(X_i)\Omega(X_i)^{-1} B_{\Omega ni}$ *and* $V_{ni}^{\dagger} = V_{ni} - D(X_i)\Omega(X_i)^{-1} V_{\Omega ni}$.

LEMMA 5. *There exist random matrices* H_j, $j = 1, \ldots, 5$, *with* $E\left[\|H_j\|^v\right] < \infty$ *for* $v = 1, 2, \ldots,$ *and* $j = 1, \ldots, 5$, *such that*

$$
\frac{\partial \widehat{s}}{\partial \theta'}(\theta_0) - J = \sum_{j=1}^{5} H_j + o_D(n^{-2\epsilon}),
$$

where

$$
\begin{aligned}
H_1 &= \frac{1}{n} \sum_{i=1}^{n} \left[D(X_i)\Omega(X_i)^{-1} D(X_i)' - J \right] + \frac{1}{n} \sum_{i=1}^{n} D(X_i)\Omega(X_i)^{-1} \eta_i' = O_p(n^{-1/2}) \\
H_2 &= \frac{1}{n} \sum_{i=1}^{n} V_{ni}^{\dagger} \Omega(X_i)^{-1} D(X_i)' = O_p(n^{-1/2}) \\
H_3 &= -\frac{1}{n} \sum_{i=1}^{n} D(X_i)\Omega(X_i)^{-1} \left\{ \frac{1}{n} \sum_{l=1}^{n} \Psi(Z_l) \right\} \Omega(X_i)^{-1} D(X_i)' = O_p(n^{-1/2}) \\
H_4 &= \sum_{\gamma=1}^{p} \frac{1}{n} \sum_{j=1}^{n} \psi_{\gamma}(Z_j) \left\{ \frac{1}{n} \sum_{i=1}^{n} F_{\gamma}(X_i)\Omega(X_i)^{-1} D(X_i)' \right\} = O_p(n^{-1/2}) \\
H_5 &= \frac{1}{n} \sum_{i=1}^{n} B_{ni}^{\dagger} \Omega(X_i)^{-1} D(X_i)' = O_p(\overline{h}^q).
\end{aligned}
$$

Furthermore, $E(H_j) = 0$, $j = 1, 2,$ *and are asymptotically normal.*

LEMMA 6. *For* $\alpha, \beta, \gamma, \delta = 1, \ldots, p$, *we have for some* $k < \infty$:

$$
\Pr\left[n^{1/2} \left| \widehat{s}_{\alpha}(\theta_0) \right| > k(\log n)^{1/2} \right] = o(n^{-2\epsilon}) \tag{36}
$$

$$
\Pr\left[n^{\epsilon} \left| \frac{\partial \widehat{s}_{\alpha}}{\partial \theta_{\beta}}(\theta_0) - E\left\{ \frac{\partial s_{\alpha}}{\partial \theta_{\beta}}(\theta_0) \right\} \right| > k(\log n)^{1/2} \right] = o(n^{-2\epsilon}) \tag{37}
$$

$$\Pr\left[n^{\epsilon}\left|\frac{\partial^2\widehat{s}_{\alpha}}{\partial\theta_{\beta}\partial\theta_{\gamma}}(\theta_0)-E\left\{\frac{\partial^2 s_{\alpha}}{\partial\theta_{\beta}\partial\theta_{\gamma}}(\theta_0)\right\}\right|>k(\log n)^{1/2}\right] = o(n^{-2\epsilon}) \quad (38)$$

$$\Pr\left[\sup_{n^{1/2}||\theta-\theta_0||\leq k\log n}\left|\frac{\partial^3\widehat{s}_{\alpha}}{\partial\theta_{\beta}\partial\theta_{\gamma}\partial\theta_{\delta}}(\theta)\right|>k(\log n)^{1/2}\right] = o(n^{-2\epsilon}). \quad (39)$$

B Appendix

We use Greek subscripts to denote either an element of a vector or partial differentiation with respect to the parameters, thus $\widehat{s}_{\alpha\beta}(\theta)=\partial\widehat{s}_{\alpha}(\theta)/\partial\theta_{\beta}$, $\widehat{s}_{\alpha\beta\gamma}(\theta)=\partial^2\widehat{s}_{\alpha}(\theta)/\partial\theta_{\beta}\partial\theta_{\gamma}$, etc. Let $s(\theta)=n^{-1}\sum_{i=1}^{n}D(X_i;\theta)\Omega(X_i)^{-1}\rho(Z_i,\theta)$, where $s(\theta)=(s_1(\theta),\ldots,s_p(\theta))'$ and $\widehat{s}(\theta)=(\widehat{s}_1(\theta),\ldots,\widehat{s}_p(\theta))'$. Let also $J_{\beta\gamma}=E\{s_{\beta\gamma}(\theta_0)\}$, $J_{\beta\delta\pi}=E\{s_{\beta\delta\pi}(\theta_0)\}$, and $(J^{\beta\gamma})=(J_{\beta\gamma})^{-1}$, and define the standardized random variables $\mathcal{Z}_{\alpha}=\sqrt{n}s_{\alpha}(\theta_0)$, $\mathcal{Z}_{\alpha\beta}=\sqrt{n}\{s_{\alpha\beta}(\theta_0)-J_{\alpha\beta}\}$, and $\mathcal{Z}_{\alpha\beta\gamma}=\sqrt{n}\{s_{\alpha\beta\gamma}(\theta_0)-J_{\alpha\beta}\}$, while $\widehat{\mathcal{Z}}_{\alpha}=\sqrt{n}\widehat{s}_{\alpha}(\theta_0)$, $\widehat{\mathcal{Z}}_{\alpha\beta}=n^{\epsilon}\{\widehat{s}_{\alpha\beta}(\theta_0)-J_{\alpha\beta}\}$, and $\widehat{\mathcal{Z}}_{\alpha\beta\gamma}=n^{\epsilon}\{\widehat{s}_{\alpha\beta\gamma}(\theta_0)-J_{\alpha\beta\gamma}\}$. Stack $J^{\bullet\beta}=(J^{1\beta},\ldots,J^{p\beta})'$ and $\widehat{\mathcal{Z}}_{\bullet}=(\widehat{\mathcal{Z}}_1,\ldots,\widehat{\mathcal{Z}}_p)'$.

In the sequel we shall make use of the facts that

$$\begin{aligned} o_D(n^{-\alpha})+o_D(n^{-\beta}) &= o_D\left(\max\left\{n^{-\alpha},n^{-\beta}\right\}\right) \quad \text{and} \\ o_D(n^{-\alpha})\cdot o_D(n^{-\beta}) &= o_D\left(n^{-(\alpha+\beta)},\max\left\{n^{-\alpha},n^{-\beta}\right\}\right). \end{aligned} \quad (40)$$

The first result is obvious. The second result is a consequence of the following argument. Let X_n be an $o_D(n^{-\alpha})$ sequence and let Y_n be an $o_D(n^{-\beta})$ sequence. Then,

$$\begin{aligned} \Pr\left[\|X_nY_n\|\geq\frac{c}{n^{(\alpha+\beta)}(\log n)^r}\right] &= \Pr\left[\left\|(n^{\alpha}X_n)(n^{\beta}Y_n)\right\|\geq\frac{c}{(\log n)^r}\right] \\ &\leq \Pr\left[\|X_n\|\geq\frac{c}{n^{\alpha}(\log n)^r}\right]+\Pr\left[\|Y_n\|\geq\frac{c}{n^{\beta}(\log n)^r}\right], \end{aligned}$$

where the inequality uses the fact that $\Pr[A\cup B]\leq\Pr[A]+\Pr[B]$ for any events A,B, and the fact that If $\|xy\|\geq a$ for $a<1$, then either $\|x\|\geq a$ or $\|y\|\geq a$.

Proof of Theorem 1. By a Taylor expansion (for $\alpha=1,\ldots,p$),

$$0 = \widehat{s}_{\alpha}(\widehat{\theta}) \quad (41)$$

$$= \widehat{s}_\alpha(\theta_0) + \sum_{\beta=1}^{p} \widehat{s}_{\alpha\beta}(\theta_0)(\widehat{\theta}_\beta - \theta_{0\beta}) + \frac{1}{2}\sum_{\beta,\gamma=1}^{p} \widehat{s}_{\alpha\beta\gamma}(\theta_0)(\widehat{\theta}_\beta - \theta_{0\beta})(\widehat{\theta}_\gamma - \theta_{0\gamma}) \quad (42)$$

$$+\frac{1}{3!}\sum_{\beta,\gamma,\delta=1}^{p} \widehat{s}_{\alpha\beta\gamma\delta}(\theta^*)(\widehat{\theta}_\beta - \theta_{0\beta})(\widehat{\theta}_\gamma - \theta_{0\gamma})(\widehat{\theta}_\delta - \theta_{0\delta}), \quad (43)$$

where θ^* are intermediate values between θ and θ_0. We first establish (11). We can rewrite (41-43) as the following system of equations:

$$-\Delta = \sum_{\beta=1}^{p} J^{\bullet\beta}\widehat{\mathcal{Z}}_\beta + \frac{1}{n^\epsilon}\sum_{\beta=1}^{p} J^{\bullet\beta}\widehat{\mathcal{Z}}_{\alpha\beta}\Delta_\beta + \frac{1}{2\sqrt{n}}\sum_{\beta,\gamma=1}^{p} J^{\bullet\beta}J_{\alpha\beta\gamma}\Delta_\beta\Delta_\gamma \quad (44)$$

$$+\frac{1}{2n^{1/2+\epsilon}}\sum_{\beta,\gamma=1}^{p} J^{\bullet\beta}\widehat{\mathcal{Z}}_{\alpha\beta\gamma}\Delta_\beta\Delta_\gamma + \frac{1}{3!n}\sum_{\beta,\gamma,\delta=1}^{p} J^{\bullet\beta}\widehat{s}_{\alpha\beta\gamma\delta}(\theta^*)\Delta_\beta\Delta_\gamma\Delta_\delta,$$

where $\Delta = n^{1/2}(\widehat{\theta} - \theta_0) = (\Delta_1, \ldots, \Delta_p)'$. Take k_0 to be the maximum of the k given in (36)-(39) [over all $\alpha, \beta, \gamma, \delta$] multiplied by $p^4 \times \|J^{-1}\|$. Then when $\|\Delta\| \leq k_0(\log n)^{1/2}$, the norm of the right hand side of (44) is less than $k_0(\log n)^{1/2}$ with probability $1 - o(n^{-2\epsilon})$, by Lemma 6. For example, the leading term satisfies

$$\{||J^{-1}\widehat{\mathcal{Z}}_\bullet|| > k_0(\log n)^{1/2}\} \subseteq \{\left\|J^{-1}\right\| \times ||\widehat{\mathcal{Z}}_\bullet|| > k_0(\log n)^{1/2} = \{||\widehat{\mathcal{Z}}_\bullet|| > \frac{k_0}{\|J^{-1}\|}(\log n)^{1/2}\},$$

which is $o(n^{-2\epsilon})$. The higher order terms are of smaller order and automatically satisfy the requirements. By Brouwer's fixed point theorem there exists a solution to (41) in the disk $||n^{1/2}(\widehat{\theta} - \theta_0)|| \leq k_0(\log n)^{1/2}$, which concludes the proof of (11). In the sequel we shall suppose that $\widehat{\theta}$ is any such solution and indeed can confine ourselves to the event that $||n^{1/2}(\widehat{\theta} - \theta_0)|| \leq k_0(\log n)^{1/2}$.

We now establish (12). Let $\widehat{\theta}^* = (\widehat{\theta}_1^*, \ldots, \widehat{\theta}_p^*)'$ solve the truncated equations determined by letting the right hand side of (42) equal zero. Collect the array $\{\widehat{s}_\alpha(\theta_0), \widehat{s}_{\alpha\beta}(\theta_0), \widehat{s}_{\alpha\beta\gamma}(\theta_0)\}_{\alpha,\beta,\gamma=1}^{p}$ in a single vector $\widehat{\mathcal{Z}}$ of dimensions $p^\dagger = \sum_{r=1}^{3}\binom{p+r-1}{r}$, and let μ denote its probability limit. Consider the p equations

$$z_\alpha + \sum_{\beta=1}^{p} z_{\alpha\beta}\delta_\beta + \frac{1}{2}\sum_{\beta,\gamma=1}^{p} z_{\alpha\beta\gamma}\delta_\beta\delta_\gamma = 0, \qquad 1 \leq \alpha \leq p \quad (45)$$

in the $p^\dagger + p$ variables $z = \{z_\alpha, z_{\alpha\beta}, z_{\alpha\beta\gamma}\}_{\alpha,\beta,\gamma=1}^{p}$ and $\delta = \{\delta_\beta\}_{\beta=1}^{p}$. These have a solution at $\delta = 0$ and $z = \mu$, i.e., $z_\alpha = 0$, $\alpha = 1, \ldots, p$. By the implicit function theorem, there is a neighborhood $\mathcal{N}$ of μ and an infinitely differentiable vector of functions $H = (H_1, \ldots, H_p)'$ such that $\delta = H(z)$, satisfies (45) for z in $\mathcal{N}$, where $H(\mu) = 0$. We are actually confined already to a shrinking neighborhood of zero by virtue of (11). Now take $z = \widehat{\mathcal{Z}}$ and $z = \widehat{\mathcal{Z}}^*$, where $\widehat{\mathcal{Z}}^*$ is the same as $\widehat{\mathcal{Z}}$ except that its first components are $\widehat{s}_\alpha(\theta_0) + \frac{1}{3!}\sum_{\beta,\gamma,\delta=1}^{p} \widehat{s}_{\alpha\beta\gamma\delta}(\theta^*)(\widehat{\theta}_\beta - \theta_{0\beta})(\widehat{\theta}_\gamma - \theta_{0\gamma})(\widehat{\theta}_\delta - \theta_{0\delta})$ instead of $\widehat{s}_\alpha(\theta_0)$,

$\alpha = 1, \ldots, p$. By Lemma 6, the norm of the difference between $\widehat{\mathcal{Z}}$ and $\widehat{\mathcal{Z}}^*$ is less than $k_0(\log n)^{1/2} n^{-1/2}$ with probability $1 - o(n^{-2\epsilon})$ when $||n^{1/2}(\widehat{\theta} - \theta_0)|| \leq k_0(\log n)^{1/2}$. Therefore, by the uniqueness part of the implicit function theorem we have $\widehat{\theta}^* - \theta_0 = H(\widehat{\mathcal{Z}}^*)$ and $\widehat{\theta} - \theta_0 = H(\widehat{\mathcal{Z}})$ with probability $1 - o(n^{-2\epsilon})$. Let $T_n = \sqrt{n}(\widehat{\theta} - \theta_0)$ and $T_n^* = \sqrt{n}(\widehat{\theta}^* - \theta_0)$. The discrepancy between $\widehat{\theta}$ and $\widehat{\theta}^*$ can be shown to be small by a Taylor expansion. Indeed,

$$\Pr\left[n^{2\epsilon} \left\|T_n - T_n^*\right\| > k(\log n)^{1/2}\right] = o(n^{-2\epsilon}). \tag{46}$$

Now let

$$\begin{aligned} T_n^{**} &= -\sum_{\beta=1}^{p} J^{\bullet\beta} \widehat{\mathcal{Z}}_\beta + \frac{1}{n^\epsilon} \sum_{\beta,\gamma,\delta=1}^{p} J^{\bullet\beta} J^{\gamma\delta} \widehat{\mathcal{Z}}_{\beta\gamma} \widehat{\mathcal{Z}}_\delta - \frac{1}{2\sqrt{n}} \sum_{\beta,\gamma,\delta,\lambda,\pi=1}^{p} J^{\bullet\beta} J^{\gamma\delta} J^{\lambda\pi} J_{\beta\delta\pi} \widehat{\mathcal{Z}}_\gamma \widehat{\mathcal{Z}}_\lambda \\ &\quad - \frac{1}{2n^{2\epsilon}} \sum_{\beta,\gamma,\delta,\pi,\nu=1}^{p} J^{\bullet\beta} J^{\gamma\pi} J^{\nu\delta} \widehat{\mathcal{Z}}_{\beta\gamma} \widehat{\mathcal{Z}}_{\pi\nu} \widehat{\mathcal{Z}}_\delta. \end{aligned}$$

By Taylor expansion and application of Lemma 6, the random variable T_n^{**} satisfies

$$\Pr\left[n^{2\epsilon} \left\|T_n^{**} - T_n^*\right\| > k \log n\right] = o(n^{-2\epsilon}).$$

The properties of the truncated expansion T_n^{**} can now be found from the properties of the standardized random variables $\widehat{\mathcal{Z}}_\beta$, $\widehat{\mathcal{Z}}_{\beta\gamma}$, and probability limits $J_{\beta\gamma}$ and $J_{\beta\delta\pi}$. Substituting the expansions we found in Lemmas 4-6 in Appendix A, we obtain

$$T_n^{**} = \sum_{\ell=0}^{5} T_{n\ell} + o_D(n^{-2\epsilon}) = T_n^{***} + o_D(n^{-2\epsilon}), \tag{47}$$

where the leading term $T_{n0} = -\sum_{\beta=1}^{p} J^{\bullet\beta} S_{0\beta} = O_p(1)$, while

$$\begin{aligned} T_{n1} &= \sum_{\beta,\gamma,\delta=1}^{p} J^{\bullet\beta} J^{\gamma\delta} H_{3\beta\gamma} \mathcal{Z}_\delta - \sum_{\beta=1}^{p} J^{\bullet\beta} S_{3\beta} \quad ; \quad T_{n2} = -\sum_{\alpha,\beta=1}^{p} J^{\bullet\beta} S_{4\beta} \\ T_{n3} &= -\sum_{\beta=1}^{p} J^{\bullet\beta}(S_{1\beta} + S_{2\beta}) + \sum_{\beta,\gamma,\delta=1}^{p} J^{\bullet\beta} J^{\gamma\delta} \left(H_{1\beta\gamma} + H_{2\beta\gamma}\right) \mathcal{Z}_\delta - \frac{1}{\sqrt{n}} \sum_{\beta,\gamma,\delta,\lambda,\pi=1}^{p} J^{\bullet\beta} J^{\gamma\delta} J^{\lambda\pi} J_{\beta\delta\pi} \mathcal{Z}_\gamma \mathcal{Z}_\lambda \\ T_{n4} &= \sum_{\beta,\gamma,\delta=1}^{p} J^{\bullet\beta} J^{\gamma\delta} H_{3\beta\gamma} S_{3\delta} - \frac{1}{2} \sum_{\beta,\gamma,\delta,\pi,\nu=1}^{p} J^{\bullet\beta} J^{\gamma\pi} J^{\nu\delta} H_{3\beta\gamma} H_{3\pi\nu} \mathcal{Z}_\delta \\ T_{n5} &= \sum_{\beta,\gamma,\delta=1}^{p} J^{\bullet\beta} J^{\gamma\delta} H_{3\beta\gamma} S_{4\delta}. \end{aligned}$$

Here, $T_{n1} = O_p(\overline{h}^q)$, $T_{n2} = O_p(1/\sqrt{n\underline{h}^d})$, $T_{n3} = O_p(n^{-1/2})$, $T_{n4} = O_p(\overline{h}^{2q})$, and $T_{n5} = O_p(\overline{h}^q/\sqrt{n\underline{h}^d})$. Write $T_{na} = T_{n0} + T_{n1} + T_{n4}$, $T_{nb} = T_{n3}$, and $T_{nc} = T_{n2} + T_{n5}$. The quantities μ_n, Ψ_n in Theorem 1 are the mean and variance of T_n^{***}.

Define the random variables

$$U_1 = \frac{1}{\sqrt{n}} \sum_{i=1}^{n} \zeta_n(Z_i) \tag{48}$$

$$U_2 = \frac{1}{\sqrt{n}} \sum_{j=1}^{n} \sum_{\substack{i=1 \\ i \neq j}}^{n} w_{ij} \varphi_n(Z_i, Z_j), \tag{49}$$

with $\zeta_n(\cdot)$ a deterministic vector function satisfying $E[\zeta_n(Z_i)] = 0$ and $\sup_n E[\|\zeta_n(Z_i)\|^r] < \infty$, for $r = 1, 2, \ldots$, while $\varphi_n(\cdot, \cdot)$ is a deterministic vector function satisfying $E[\varphi_n(Z_i, Z_j)|Z_i] = E[\varphi_n(Z_i, Z_j)|Z_j] = 0$ with probability one, and $\sup_n E[\|\varphi_n(Z_i, Z_j)\|^r] < \infty$, $r = 1, 2, \ldots$. Note that T_{na} is of the form (48), the random sequence T_{nb} is a homogenous quadratic polynomial in such sums with magnitude $O_p(n^{-1/2})$, while T_{nc} is like (49) with $\varphi_n(Z_i, Z_j) = \sum_{\beta,\gamma,\delta=1}^{p} \{J^{\bullet\delta} + J^{\bullet\beta} J^{\gamma\delta} H_{3\beta\gamma}\} (\eta_j^\dagger \Omega^{-1} \rho_i)_\delta$.

We now establish the validity of the Edgeworth approximation to the distribution of T_n^{***} from which follows the validity of the distributional approximation for T_n. Our method of proof extends the univariate results of Nishiyama and Robinson (1997) to higher dimensions and smaller order of error [they considered only $O(n^{-1/2})$ approximations]. The random sequence $T_{na} + T_{nb}$ has a valid Edgeworth expansion to $o(n^{-2\epsilon})$, as follows from standard results for parametric estimators such as can be found in Bhattacharya and Ghosh (1979), the main problem arises with T_{nc} and its interaction with T_{na}. For expositional reasons we shall therefore just establish the result for $T_{na} + T_{nc}$, since incorporating T_{nb} is conceptually trivial but notationally complicated. By the so-called smoothing lemma [Bhattacharya and Rao (1976, Lemma 12.1 and Lemma 12.2) and Götze (1987, 3.4)], the left hand side of (12) is bounded by

$$\max_{|\mathbf{a}| \leq p+1} \int_0^{n^{2\epsilon} \log n} \left| \partial^{\mathbf{a}} \{\psi_n(s) - \widetilde{\psi}_n(s)\} \right| ds + o(n^{-2\epsilon}) \tag{50}$$

for some constant k, where $\psi_n(s)$ is the characteristic function of T_n^{***}, while $\widetilde{\psi}_n(s)$ is the Fourier transform of the signed measure $\widetilde{F}_{n1}^*(x)$ [which is $\widetilde{F}_{n1}(x)$ renormalized to have mean vector μ_n and variance matrix Ψ_n]. Here, $\int_a^b \cdot ds$ denotes integration over the p-dimensional region where $a < \|s\| < b$. By the triangle inequality,

$$\int_0^{n^{2\epsilon} \log n} \left| \partial^{\mathbf{a}} \{\psi_n(s) - \widetilde{\psi}_n(s)\} \right| ds \leq \int_0^{n^\delta} \left| \partial^{\mathbf{a}} \{\psi_n(s) - \widetilde{\psi}_n(s)\} \right| ds + \int_{n^\delta}^{n^{2\epsilon} \log n} |\partial^{\mathbf{a}} \psi_n(s)| \, ds + \int_{\log n}^{\infty} \left| \partial^{\mathbf{a}} \widetilde{\psi}_n(s) \right| ds$$

for any $\delta > 0$. We take $\delta = (1 - 2\epsilon)/(p + 5)$. The last integral is $o(n^{-2\epsilon})$ because of the form of the Edgeworth characteristic function $\widetilde{\psi}_n$.

We must develop some machinery for dealing with $\partial^{\mathbf{a}}\psi_n(s)$ and $\partial^{\mathbf{a}}\widetilde{\psi}_n(s)$. For any set $L_m = \{i_1, \ldots, i_m\}$ with $m = m(n)$, let $\Delta'_n(m) = \sum\sum_{i,j\notin L_m} w_{ij}\varphi_n(Z_i, Z_j)$ and $\Delta_n(m) = T_{nc} - \Delta'_n(m)$. Then, $T_n^{***} - \Delta_n(m) = \sum_{i\in L_m}\zeta_n(Z_i)/\sqrt{n} + \Delta_L(m)$, where $\Delta_L(m)$ does not depend on Z_j, $j \in L_m$. When $m = n$, $\Delta_n(m) = T_{nc}$ and $\Delta_L(m) \equiv 0$. Note that under our moment conditions, $E\left[\|\Delta_n(m)\|^r\right] = O(m^{r/2}/n^{r(1+2\epsilon)/2})$ for any integer r. By Taylor expansion we have for any integer r,

$$\psi_n(s) = \overline{\psi}_{n,m,r}(s) + R_{m,r}(s) \equiv \sum_{\ell=0}^{r}\psi_{n,m,\ell}(s) + R_{m,r}(s), \tag{51}$$

where $\psi_{n,m,\ell}(s) = E\left[\exp(\mathbf{i}s'(T_n^{***} - \Delta_n(m)))\left(is'\Delta_n(m)\right)^{\ell}/\ell!\right]$ and

$$R_{m,r}(s) = E\left[e^{is'(T_n^{***} - \Delta_n(m)\cdot u)}\frac{\left(is'\Delta_n(m)\right)^{(r+1)}}{(r+1)!}\right],$$

where u is uniformly distributed on $[0,1]$ independently of the data, while

$$\psi_{n,m,\ell}(s) = \frac{1}{\ell! n^{\ell/2}}\sum{}^{*}\xi_n^{m-(J\cap L_m)_{\#}}(s)\cdot w_{i_1 j_1}\cdots w_{i_\ell j_\ell}\cdot\Gamma_{n\ell}(s; J, m),$$

where the summation is over all 2ℓ-tuples $J = (i_1, j_1, \ldots, i_\ell, j_\ell)$ with $i_s \neq j_s$ and $A_{\#}$ is the number of distinct elements in the set A, while

$$\Gamma_{n\ell}(s; J, m) = E\left[\exp\left(\frac{i}{\sqrt{n}}\sum_{j\in J\cap L_m} s'\zeta_n(Z_j)\right)\exp(\mathbf{i}s'\Delta_L(m))s'\varphi_n(Z_{i_1}, Z_{j_1})\cdots s'\varphi_n(Z_{i_\ell}, Z_{j_\ell})\right].$$

The analysis of $\xi_n^{m-(J\cap L_m)_{\#}}(s)$ and $\Gamma_{n\ell}(s; J, n)$ and their partial derivatives follows from similar work to Götze (1987); the behaviour of nonparametric weights follows from Lemma 3.

We shall use the following fact: for any p-vectors $x = (x_1, \ldots, x_p)'$ and $\mathbf{b} = (b_1, \ldots, b_p)'$ and integer r, we have $\partial^{\mathbf{b}}e^{is'x} = i^{|\mathbf{b}|}e^{is'x}\prod_{j=1}^{p}x_j^{b_j}$ and $\partial^{\mathbf{b}}(s'x)^r = (r)_{|\mathbf{b}|}(s'x)^{r-|\mathbf{b}|}\prod_{j=1}^{p}x_j^{b_j}$, where $(r)_{|\mathbf{b}|} = r\cdots(r-|\mathbf{b}|)$. We first take $m = n$. Then for any vector $\mathbf{a}$ with $|\mathbf{a}| \leq p+1$ we have by the chain rule and the fact that $|\exp(\mathbf{i}x)| \leq 1$ for all real x,

$$\begin{aligned}
|\partial^{\mathbf{a}}R_{n,r}(s)| &\leq k\cdot\sum_{\mathbf{b}+\mathbf{c}=\mathbf{a}}E\left[\left|\partial^{\mathbf{b}}e^{is'(T_{na}+T_{nc}\cdot u)}\cdot\partial^{\mathbf{c}}(s'T_{nc})^{r+1}\right|\right] \\
&\leq k\cdot\sum_{\mathbf{b}+\mathbf{c}=\mathbf{a}}E\left[\left|(s'T_{nc})^{r+1-|\mathbf{b}|}\cdot\prod_{j=1}^{p}(T_{na}+T_{nc}\cdot u)_j^{b_j}(T_{nc})_j^{c_j}\right|\right] \\
&\leq k\cdot\left(1+\|s\|^{r+1}\right)E\left[\|T_{nc}\|^{r+1}\left(1+\|T_{nc}\|^{p+1}\right)\|T_{na}\|^{p+1}\right],
\end{aligned}$$

by the Cauchy-Schwarz inequality. Then, since $\int_0^{n^\delta}\left(1+\|s\|^{r+1}\right)ds = O(n^{\delta(p+r+2)})$ and

$$E\left[\|T_{nc}\|^{r+1}\left(1+\|T_{nc}\|^{p+1}\right)\|T_{na}\|^{p+1}\right] = O(n^{-(r+1)\epsilon}),$$

we have shown that

$$\max_{|\mathbf{a}|\leq p+1} \int_0^{n^\delta} |\partial^{\mathbf{a}} R_{n,r}(s)|\, ds = o(n^{-2\epsilon}),$$

provided $r > \{\epsilon + \delta(p+2)\}/(\epsilon - \delta)$.

We next deal with the term $\int_0^{n^\delta} \left|\sum_{\ell=3}^r \partial^{\mathbf{a}} \psi_{n,n,\ell}(s)\right| ds$. First note that

$$\left|\partial^{\mathbf{a}} \psi_{n,n,r}(s)\right| \leq k \cdot \sum_{\mathbf{b}+\mathbf{c}=\mathbf{a}} \left|\partial^{\mathbf{b}} \xi_n^{m-(J\cap L_m)_\#}(s)\right| \frac{1}{n^{\ell/2}} \sum{}^* |w_{i_1 j_1} \cdots w_{i_\ell j_\ell}| \cdot |\partial^{\mathbf{c}} \Gamma_{nr}(s; J, n)|\,,$$

by crude bounding. Furthermore, there is some polynomial $P(\cdot)$ with bounded coefficients and constant $k > 0$ such that

$$\left|\partial^{\mathbf{b}} \xi_n^{m-(J\cap L_m)_\#}(s)\right| \leq \exp(-k\,\|s\|^2) P(\|s\|),$$

by Götze (1987, Lemma 3.3). Therefore, it suffices to estimate $\sum^* |w_{i_1 j_1} \cdots w_{i_\ell j_\ell}| \cdot |\partial^{\mathbf{c}} \Gamma_{nr}(s; J, n)|$. Consider the case $\mathbf{a} = \mathbf{c} = 0$ and $\ell = 3$, in which situation there are five subcases $J_\# = 2, \ldots, J_\# = 6$. By Lemma 3,

$$\begin{aligned} \sum{}^* w_{ij}^3 &= O(n^{1-4\epsilon}),\ \sum{}^* |w_{ij} w_{jk} w_{ki}| = O(n^{1-2\epsilon}),\ \sum{}^* |w_{ij} w_{jk} w_{kl}| = O(n^{2-2\epsilon}), \quad (52) \\ \sum{}^* |w_{ij} w_{jk} w_{rs}| &= O(n^{3-2\epsilon}),\ \sum{}^* |w_{ij} w_{kl} w_{rs}| = O(n^3), \end{aligned}$$

where the sums are over two, three, four, five, and six distinct indices respectively. Now let $\zeta_{nj}(s) = \exp(\mathbf{i} s' \zeta_n(Z_j)/\sqrt{n})$. We have on the range $0 \leq \|s\| \leq o(\sqrt{n})$ that:

$$\begin{aligned} E\left[\zeta_{ni}(s)\zeta_{nj}(s) s'\varphi_n(Z_i, Z_j)\right] &= O(\|s\|^3/n), \quad (53) \\ E\left[\prod_\ell \zeta_{n\ell}(s)\{s'\varphi_n(Z_i, Z_j)\}^2 s'\varphi_n(Z_j, Z_k)\right] &= O(\|s\|^4/\sqrt{n}), \\ E\left[\prod_\ell \zeta_{n\ell}(s) s'\varphi_n(Z_i, Z_j) s'\varphi_n(Z_j, Z_k) s'\varphi_n(Z_k, Z_i)\right] &= O(\|s\|^3), \\ E\left[\prod_\ell \zeta_{n\ell}(s) s'\varphi_n(Z_i, Z_j) s'\varphi_n(Z_j, Z_k) s'\varphi_n(Z_k, Z_l)\right] &= O(\|s\|^5/n), \\ E\left[\prod_\ell \zeta_{n\ell}(s) s'\varphi_n(Z_i, Z_j) s'\varphi_n(Z_j, Z_k) s'\varphi_n(Z_l, Z_r)\right] &= O(\|s\|^6/n\sqrt{n}) \\ E\left[\prod_\ell \zeta_{n\ell}(s) s'\varphi_n(Z_i, Z_j) s'\varphi_n(Z_k, Z_l) s'\varphi_n(Z_r, Z_t)\right] &= O(\|s\|^9/n^3). \quad (54) \end{aligned}$$

The magnitude of the expectations (53)-(54) is established as follows. Write $\zeta_{ni}(s) = P_q(s'\zeta_n(Z_i)/\sqrt{n}) + R_q(s'\zeta_n(Z_i)/\sqrt{n})$, where $P_q(x) = 1 + ix + \ldots + (ix)^q/q!$ is the q^{th} order expansion of $\exp(\mathbf{i}x)$ about

$x = 0$ and $R_q(x) = \exp(\mathbf{i}x) - P_q(x)$. Then, multiply out and take expectation term by term [the polynomial terms yield zeros when there is an odd Z_j]. In conclusion, we obtain $\left|\psi_{n,n,r}(s)\right| \leq \exp(-k\left\|s\right\|^2)P_*(\|s\|)o(n^{-2\epsilon})$ for some polynomial P_* with bounded coefficients. Thus, the integral over $0 \leq \|s\| \leq n^\delta$ is $o(n^{-2\epsilon})$ as required. The same argument applies to general $\mathbf{a}$, because: (a) differentiating the polynomial terms with respect to s simply changes the polynomial P_* in our bound, while (b) as for the remainder terms, we have for any integer q and any vector $\mathbf{a}$, $|\partial^{\mathbf{a}} R_q(s'\zeta_n(Z_i)/\sqrt{n}) \leq k|R_q(s'\zeta_n(Z_i)/\sqrt{n})|$ for some constant k. Finally, under our conditions, $\left|\partial^{\mathbf{a}}\psi_{n,n,\ell}(s)\right| \leq k\left|\partial^{\mathbf{a}}\psi_{n,n,\ell+1}(s)\right|$ for any ℓ. Therefore, combining (52) and (53)-(54), we get

$$\int_0^{n^\delta} \left|\sum_{\ell=3}^{r} \partial^{\mathbf{a}}\psi_{n,n,\ell}(s)\right| ds = o(n^{-2\epsilon})$$

as required.

We next establish that

$$\int_{n^\delta}^{n^{2\epsilon}\log n} |\partial^{\mathbf{a}}\psi_n(s)|\, ds = o(n^{-2\epsilon}). \tag{55}$$

For this we use the expansion (51) with $m(n) < n$. We just consider the case $\mathbf{a} = 0$, the general case follows similarly. We have

$$|\psi_n(s)| \leq k\left|\xi_n^{m-2\ell}(s)\right| \cdot \sum_{\ell=0}^{r} \|s\|^\ell \frac{m^\ell}{n^{\ell(1+2\epsilon)}} + k \cdot \|s\|^{r+1} \frac{m^{r/2}}{n^{r(1+2\epsilon)/2}}, \tag{56}$$

by an extension of some results of Callaert, Janssen, and Verarbereke (1984) to the multivariate case with kernel weighting [see also Bickel, Götze, and van Zwet (1986, 2.17)]. Without loss of generality we suppose that $\epsilon > 1/4-$ when $\epsilon \leq 1/4$ this means that $2\epsilon \leq 1/2$ and the argument below is even simpler. We first consider the range $n^\delta \leq \|s\| \leq k\sqrt{n}$ for some constant k, which we split up into subintervals $I_1 = n^\delta \leq \|s\| \leq n^\epsilon$ and $I_2 = n^\epsilon \leq \|s\| \leq k\sqrt{n}$. On each interval we take $m(n,s) = O(n^{\alpha^j})$, where $\alpha^1 > \alpha^2 > 0$. On I_1, the first term in (56) is $o(n^{-2\epsilon})$, because $|\xi_n(s)| \leq \exp(-k\left\|s\right\|^2/n)$ when $\|s\| \leq k\sqrt{n}$, and $\left|\xi_n^{m-2\ell}(s)\right| \leq k\exp(-kn^{\alpha^1+2\delta-1}) = o(n^{-d})$ for any d, provided $n^\delta < \|s\|$ and $\alpha^1 > 1 - 2\delta$. The second term in (56) contributes $o(n^{-2\epsilon})$ provided $r > \epsilon(8+2p)/(1-\alpha^1)$, because $\int_{n^\delta}^{n^{\delta^1}} \|s\|^{r+1} ds = O(n^{(p+r+2)\delta^1})$ for any δ^1. On I_2, we must take $\alpha^2 > 1 - 2\epsilon$ and $r > (4\epsilon + 2 + p)/(2\epsilon - \alpha^2)$.

On the range $k\sqrt{n} \leq \|s\| \leq n^{2\epsilon}\log n$ we have $|\xi_n(s)| \leq 1 - \vartheta$ for some $\vartheta > 0$. Take $m(n) = O(\log n)$. Then, the first term in (56) is bounded by a constant times $|1-\vartheta|^{k\log n}\sum_{\ell=0}^{r}\|s\|^\ell (\log n)^\ell n^{-\ell(1+2\epsilon)}$. For large k, this is small enough. As for the second term in (56), this is small enough because $n^{(r+2+p)2\epsilon}(\log n)^k n^{-r(1+2\epsilon)/2}$ can be made smaller than order $n^{-2\epsilon}$ by taking $r > (12+4p)\epsilon/(1-2\epsilon)$.

Finally, we show that

$$\int_0^{n^\delta} \left| \sum_{\ell=0}^{2} \partial^{\mathbf{a}} \{\psi_{n,n,\ell}(s) - \widetilde{\psi}_n(s)\} \right| ds = o(n^{-2\epsilon}). \tag{57}$$

Consider again the case $\mathbf{a} = 0$. Using the conditioning argument and the magnitudes of the weights, we obtain

$$\begin{aligned} E\left[e^{is'T_{na}} s'T_{nc}\right] &= \xi_n^n(s) \times \left[\frac{-1}{\sqrt{n}} E\left\{s'\zeta_n(Z_1)s'\zeta_n(Z_2)s'\varphi_n(Z_1,Z_2)\right\} + O\left(\frac{\|s\|^3}{n}\right)\right] \\ E\left[e^{is'T_{na}} (s'T_{nc})^2\right] &= \xi_n^n(s) \times \left[E\{(s'T_{nc})^2\} + O\left(\frac{\|s\|^4}{n}\right)\right], \end{aligned}$$

and after integrating over the range $0, n^\delta$ we obtain $o(n^{-2\epsilon})$ errors as required. The general $\mathbf{a}$ case is similar.

■

PROOF OF COROLLARY 1. This follows by a standard application of deJong (1987, Theorem 2.1 and Proposition 3.2) and of the Cramér-Wold device, and is worked out in more detail in Linton (1997). ■

PROOF OF THEOREM 2. We just give a sketch of the argument because it mostly follows from Theorem 1. By Taylor series expansion, $\sqrt{n}\widehat{g} = G_0\sqrt{n}(\widehat{\theta} - \theta_0) + O_p(n^{-1/2})$, while $\widehat{G} = G_0 + O_p(n^{-1/2})$, and $\widehat{J} = J + O_p(\overline{h}^{*q}) + O_p(n^{-1/2})$. Furthermore, we obtain

$$\widehat{W} = \sqrt{n}(\widehat{\theta} - \theta_0)' Q \sqrt{n}(\widehat{\theta} - \theta_0) + \Gamma_n + o_D(n^{-2\epsilon}), \tag{58}$$

where Γ_n is a homogeneous cubic polynomial in a vector U_n of asymptotically normal single sums of independent random variables and is $O_p(n^{-1/2})$. The expansion is given in full in Linton (1997). The validity of the Edgeworth approximation for $\widehat{W}$ now follows from the validity of a multivariate expansion for $\sqrt{n}(\widehat{\theta} - \theta_0)$ joint with U_n. This expansion follows along the same lines as Theorem 1 [the $O_p(n^{-1/2})$ terms are 'standard']. See for example Chandra and Ghosh (1979).

We next rewrite the expansion in more convenient form

$$\widehat{W} = \overline{W}_0 + \frac{A_n}{n^\epsilon} + \frac{B_n}{\sqrt{n}} + \frac{C_n}{n^{2\epsilon}} + o_D(n^{-2\epsilon}), \tag{59}$$

where $\overline{W}_0$, A_n, B_n and C_n are stochastically bounded sequences obtained from the expansion of $\sqrt{n}(\widehat{\theta} - \theta_0)$. In fact, $\overline{W}_0 = X'QX$ [which has distribution $\chi^2_{p_1}$ to order n^{-1}], $A_n = 2X'Q\Delta$, and

$C_n = \Delta' Q \Delta$, where X [$= T_{na}$] is the leading term of $\sqrt{n}(\widehat{\theta} - \theta_0)$ and Δ [$= T_{nc}$] is the second order term. We next apply the algorithm described in Rothenberg (1984), which, with suitable modification for the different orders of magnitudes we have, says that

$$\Pr\left[\widehat{W} \leq x\right] = F_{p_1,\infty}\left[x - \frac{a(x)}{n^{\epsilon}} - \frac{b(x)}{\sqrt{n}} + \frac{2a(x)a'(x) + q(x)v(x) + v'(x) - 2c(x)}{2n^{2\epsilon}}\right] + o(n^{-2\epsilon}),$$

where $q(x) = (p_1 - 2 - x)/2x = d \log F'_{p_1,\infty}(x)/dx$, while $a(x) = E\left(A_n | \overline{W}_0 = x\right)$, $b(x) = E\left(B_n | \overline{W}_0 = x\right)$, $c(x) = E\left(C_n | \overline{W}_0 = x\right)$, and $v(x) = \text{var}\left(A_n | \overline{W}_0 = x\right)$. The random variable B_n is a homogeneous cubic polynomial in asymptotically normal single sums of independent random variables, while $\overline{W}_0$ is a quadratic function of similar random variables; this implies that $b(x) = O(n^{-1/2})$ by standard arguments. Also, $a(x) = o(n^{-\epsilon})$ by the arguments used in Linton (1997, expression (54)). Finally, by the law of iterated expectation

$$\begin{aligned} v(x) &= 4E\left[X'QE\left(\Delta\Delta'|X\right)QX|X'QX = x\right] \\ &= 4E\left[X'Q\Sigma QX|X'QX = x\right] + o(1) \\ &= \frac{4\text{tr}\left(Q\Sigma\right)}{p_1} \cdot x + o(1), \end{aligned}$$

because Δ is asymptotically independent of X, and because the best prediction of $X'Q\Sigma QX$ by $X'QX$ is linear with coefficient $\text{cov}(X'Q\Sigma QX, X'QX)/\text{var}(X'QX)$. Similarly,

$$c(x) = E\left[\text{tr}\left(\Delta\Delta'Q|X\right)|X'QX = x\right] = \text{tr}\left(Q\Sigma\right).$$

■

PROOF OF THEOREM 3. We have

$$\begin{aligned} &\Pr\left[\sqrt{n}\left\|\widehat{\theta}(\widehat{\lambda}_{opt}) - \widehat{\theta}(\lambda_{opt})\right\| > \frac{k}{n^{2\epsilon}\log n}\right] \\ \leq\ &\Pr\left[\sup_{\|\lambda - \lambda_{opt}\| \leq \delta n^{-\varphi_1}} \sqrt{n}\left\|\widehat{\theta}(\lambda) - \widehat{\theta}(\lambda_{opt})\right\| > \frac{k}{n^{2\epsilon}\log n}\right] + \Pr\left[n^{\varphi_1}\left\|\widehat{\lambda} - \lambda_{opt}\right\| > \delta\right] \\ =\ &o(n^{-2\epsilon}), \end{aligned}$$

because of (18) and

$$\sup_{||\lambda - \lambda_{opt}|| \leq \delta n^{-\varphi_1}} \sqrt{n}\left\|\widehat{\theta}(\lambda) - \widehat{\theta}(\lambda_{opt})\right\| \leq \delta n^{1/2-\varphi_1} \times \sup_{||\lambda - \lambda_{opt}|| \leq \delta n^{-\varphi_1}} \left\|\frac{\partial \widehat{\theta}}{\partial \lambda}(\lambda)\right\| = o_D(n^{1/2-(\varphi_1+\varphi_2)}, n^{-2\epsilon}),$$

which follows by the Mean Value Theorem, crude bounding, and (19). ■

C Proofs of Lemmas

Our arguments below will make use of the following standard exponential inequality.

LEMMA EXP. *Let Z_{ni} be independent random variables with $E(Z_{ni}) = 0$ and $\text{var}(Z_{ni}) = \sigma_{ni}^2$, where $\sum_{i=1}^n \sigma_{ni}^2 > 0$. Suppose also that $|Z_{ni}| \leq m$ for some constant m. Then, for any $\lambda > 0$,*

$$\Pr\left[\left|\sum_{i=1}^n Z_{ni}\right| > \lambda\right] \leq 2\exp\left[-\frac{\lambda^2}{2\left(\sum_{i=1}^n \sigma_{ni}^2 + m\lambda\right)}\right].$$

PROOF OF LEMMA 1. The proof of a stronger [involving supremum] version of (25) and (27) for the case where $h_{ni} = h_n$ is given in the recent paper by Masry (1996). The extension to our case is straightforward under our assumptions A1 and A5. In particular, there is a finite constant k such that

$$\left|\sum_j w_j(x)T_j - \sum_j w_j(x')T_j\right| \leq k\underline{h}^{-(d+1)}|x - x'|$$

for any bounded sequence T_j and any points x, x'. This leads to the same proof of his 3.22, 3.23, and 4.21. ■

PROOF OF LEMMA 2. Let $v_{ni} = \left(M_{ni} - \overline{M}_{ni}\right)_{\alpha\beta} = n^{-1}h_{ni}^{-1}\sum_{j\neq i} Z_{nij}$, where

$$Z_{nij} = K\left(\frac{X_i - X_j}{h_{ni}}\right)\left(\frac{X_i - X_j}{h_{ni}}\right)^{\mathbf{a}} - E\left\{K\left(\frac{X_i - X_j}{h_{ni}}\right)\left(\frac{X_i - X_j}{h_{ni}}\right)^{\mathbf{a}}|X_i\right\}$$

are mutually independent given X_i. We apply Lemma Exp with Z_{nij} as above and $\lambda = kn^\epsilon \log n$. Since K is bounded and has bounded support, $|Z_{ni}| \leq 2\sup_u |K(u)u^{\mathbf{a}}| \equiv \overline{K}$, i.e., we can take $m(n) = \overline{K}$, while $\text{var}\left(\sum_{j\neq i} Z_{nij}\,|X_i\right) = s_{ni}n^{2\epsilon}$, where $s_{ni} \leq \overline{s}$ for some finite constant $\overline{s}$. We have with probability one

$$\begin{aligned}
\Pr\left[|v_{ni}| > k\frac{\log n}{n^\epsilon}\,|X_i\right] &\leq 2\exp\left[-\frac{k^2n^{2\epsilon}(\log n)^2}{2\left\{\left(\text{var}\left(\sum_{j\neq i} Z_{nij}\,|X_i\right) + kn^\epsilon \log n\right)\right\}}\right] \\
&= 2\exp\left[-\frac{k^2n^{2\epsilon}(\log n)^2}{2\left(s_{ni}n^{2\epsilon} + kn^\epsilon \log n\right)}\right] \\
&= o(n^{-1-2\epsilon}),
\end{aligned}$$

for large k. Therefore, by the Bonferroni inequality,

$$\Pr\left(\max_{1\leq i\leq n}|v_{ni}|>k\frac{\log n}{n^{\epsilon}}\,|X_i\right)\leq\sum_{i=1}^{n}\Pr\left[|v_{ni}|>k\frac{\log n}{n^{\epsilon}}\,|X_i\right]=o(n^{-2\epsilon})$$

with probability one, as required. The argument for (29) is a straightforward consequence of the smoothness assumptions.

The argument for (30) is similar to that for (28) except that we must use a truncation argument to compensate for the unboundedness of η_j. Define $\eta_j'=\eta_j 1(|\eta_j|\leq n^{\epsilon})$ and $\eta_j''=\eta_j\mathbf{1}(|\eta_j|>n^{\epsilon})$. We show that

$$\Pr\left[\max_{1\leq i\leq n}\left|\sum_j w_{ij}\{\eta_j'-E(\eta_j')\}\right|>k\frac{\log n}{n^{\epsilon}}\right]=o(n^{-2\epsilon}) \tag{60}$$

$$\Pr\left[\max_{1\leq i\leq n}\left|\sum_{j}^{n} w_{ij}\{\eta_j''-E(\eta_j'')\}\right|>k\frac{\log n}{n^{\epsilon}}\right]=o(n^{-2\epsilon}). \tag{61}$$

Note that by Lemma 3, $\max_{1\leq i\leq n}|w_{ij}|\leq\overline{K}n^{-2\epsilon}$ for some positive finite constant $\overline{K}$. Therefore, using Markov's inequality

$$\begin{aligned}
\Pr\left[\max_{1\leq i\leq n}\left|\sum_j w_{ij}\{\eta_j''-E(\eta_j'')\}\right|>k\frac{\log n}{n^{\epsilon}}\right] &\leq \Pr\left[\left|\sum_j n^{-2\epsilon}\left\{|\eta_j''|+E\left(|\eta_j''|\right)\right\}\right|>k\frac{\log n}{n^{\epsilon}}\right]\\
&\leq k(\log n)^{-1}n^{-2\epsilon}n^{1+\epsilon}E\left(|\eta_j''|\right)\\
&\leq k(\log n)^{-1}n^{1-\epsilon}E\left(|\eta_j|\right)\Pr\left[\mathbf{1}(|\eta_j|>n^{\epsilon})\right]\\
&\leq k(\log n)^{-1}n^{1-\epsilon}E\left(|\eta_j|\right)E\left(|\eta_j|^{\ell}\right)n^{-\ell\epsilon}\\
&= o(n^{-2\epsilon}),
\end{aligned}$$

provided $\ell\geq(1+\epsilon)/\epsilon$. Because η_j' are bounded (specifically, we have $|w_{ij}\eta_j'|\leq\overline{K}n^{-\epsilon}$), we can apply the exponential inequality. Therefore, we have, using the Bonferroni inequality,

$$\Pr\left[\max_{1\leq i\leq n}\left|\sum_j w_{ij}\{\eta_j'-E(\eta_j')\}\right|>k\frac{\log n}{n^{\epsilon}}|X_i\right]\leq\sum_{i=1}^{n}\Pr\left[\left|\sum_j w_{ij}\left\{\eta_j'-E(\eta_j')\right\}\right|>k\frac{\log n}{n^{\epsilon}}|X_i\right]$$

$$\leq \quad 2n \exp\left[-\frac{k^2 n^{-2\epsilon}(\log n)^2}{2\left\{v n^{-2\epsilon} + k\overline{K} n^{-2\epsilon}\log n\right\}}\right]$$

$$= \quad o(n^{-2\epsilon}),$$

provided $k > 2\overline{K}$.

■

PROOF OF LEMMA 3. Let

$$k_{ni} = \sharp\left\{j : \left|K\left(\frac{X_i - X_j}{h_{ni}}\right)\right| > 0\right\} \quad ; \quad k_n = \max_{1\leq i\leq n} k_{ni}.$$

It suffices to show that with probability one $k_n \leq O(n^{2\epsilon})$, as $n \to \infty$. Conditional on X_i,

$$k_{ni} = \sum_{j\neq i} \mathbf{1}\left\{\left|K\left(\frac{X_i - X_j}{h_{ni}}\right)\right| > 0\right\} = \sum_{j\neq i} Z_{nij}$$

is a Binomial random variable with parameters $n-1$ and p_{ni}, and hence has conditional mean $\mu_{ni} = E[k_{ni}|X_i] = np_{ni}$ and variance $\sigma_{ni}^2 = \text{var}[k_{ni}|X_i] = (n-1)p_{ni}q_{ni}$ with $q_{ni} = 1 - p_{ni}$. We establish that the existence of constants $\underline{c}, \overline{c}$ such that with probability one

$$\underline{c}h_n^d \leq p_{ni} \leq \overline{c}h_n^d, \tag{62}$$

which implies that $\sum_{i=1}^n \sigma_{ni}^2 > 0$ with probability one. The result (62) holds because there exists rectangles A, B of fixed dimensions with $A \subseteq \text{supp}(K) \subseteq B$ such that

$$\{X_j \in X_i + h_{ni}\cdot A\} \subseteq \left\{\left|K\left(\frac{X_i - X_j}{h_{ni}}\right)\right| > 0\right\} \subseteq \{X_i - X_j \in X_i + h_{ni}\cdot B\},$$

where $x + yA$ is the set $\{x + ya : a \in A\}$. Now

$$\Pr\left[X_j \in X_i + h_{ni}\cdot A | X_i\right] \quad = \quad \int_{X_i + h_{ni}\cdot A} f(X)dX$$

$$\geq \quad \mu(X_i + h_{ni}\cdot A \cap \mathcal{S}) \inf_{x\in\mathcal{S}} f(x),$$

where μ is Lebesgue measure. For any fixed interior point x it is clear that $\liminf_{\epsilon\to 0} \mu(x + \epsilon\cdot A \cap \mathcal{S})/\epsilon^d > 0$. Assumption A1 guarantees that it is also true for boundary points x. It is also true for any sequence of points $x_n \in \mathcal{S}$. Since $\inf_{x\in\mathcal{S}} f(x) > 0$ by assumption A1, the bound (62) is established. The same argument works for the upper bound.

We now apply the exponential inequality to $k_{ni} - E\{k_{ni}|X_i\}$, noting that each Z_{nij} is bounded by one. Therefore, for any $\delta > 0$,

$$\Pr\left[|k_{ni} - E\left\{k_{ni}\,|X_i\right\}| \geq \delta n^{2\epsilon}|X_i\right] \leq 2\exp\left[-\frac{\delta^2 n^{4\epsilon}}{2\left((n-1)p_{ni}q_{ni} + 2\delta n^{2\epsilon}\right)}\right] \leq 2\exp(-kn^{2\epsilon})$$

for some constant k. Therefore, by the Bonferroni inequality and identity of distribution for large k,

$$\begin{aligned}
\Pr\left[k_n > kn^{2\epsilon}\right] &\leq n\Pr\left[k_{ni} > kn^{2\epsilon}\right] \\
&\leq n\Pr\left[E\left\{k_{ni}\,|X_i\right\} > \frac{k}{2}n^{2\epsilon}\right] + n\Pr\left[|k_{ni} - E\left\{k_{ni}\,|X_i\right\}| > \frac{k}{2}n^{2\epsilon}\right] \\
&\leq n\Pr\left[np_{ni} > \frac{k}{2}n^{2\epsilon}\right] + kn\exp(-kn^{2\epsilon})
\end{aligned}$$

using the triangle inequality. Finally, there exists n_0, k such that for all $n > n_0$, $\Pr[np_{ni} > \frac{k}{2}n^{2\epsilon}] = 0$, which implies that $\Pr\left[k_n > kn^{2\epsilon}\right] = o(n^{-2\epsilon})$ as required.

We now turn to the proof of (32). We have

$$\begin{aligned}
|w_{ij}| &\leq \sum_{\{\mathbf{a}:|\mathbf{a}|\leq q\}} \left|w_{ij}^{\mathbf{a}}\right| \\
&\leq \frac{t}{nh_{ni}^d} \cdot \max_{\{\mathbf{a}:|\mathbf{a}|\leq q\}} \left|e_1' M_{ni}^{-1} e_{\ell(\mathbf{a})}\right| \max_{\{\mathbf{a}:|\mathbf{a}|\leq q\}} \left|K\left(\frac{X_i - X_j}{h_{ni}}\right)\left(\frac{X_i - X_j}{h_{ni}}\right)^{\mathbf{a}}\right| \\
&\leq \frac{t}{nh_{ni}^d} \cdot \max_{\{\mathbf{a}:|\mathbf{a}|\leq q\}} \left|e_1' M_{ni}^{-1} e_1\right|^{1/2} \left|e_{\ell(\mathbf{a})}' M_{ni}^{-1} e_{\ell(\mathbf{a})}\right|^{1/2} \max_{\{\mathbf{a}:|\mathbf{a}|\leq q\}} \sup_u |K(u)u^{\mathbf{a}}| \\
&\leq \frac{t}{nh_{ni}^d} \cdot \lambda_{\max}(M_{ni}^{-1}) \max_{\{\mathbf{a}:|\mathbf{a}|\leq q\}} \sup_u |K(u)u^{\mathbf{a}}|\,.
\end{aligned}$$

Finally, with probability one $\lambda_{\max}(M_{ni}^{-1}) = \lambda_{\max}(M_i^{-1}) + o(1)$, where the matrix M_i is uniformly positive definite by A1 and A4.

The proof of (33) follows similar lines and uses the fact that uniformly in i, $\sum_{j\neq i} |K((X_i - X_j)/h_{ni})((X_i - X_j)/h_{ni})^{\mathbf{a}}| = O(n\overline{h}_n^d)$ with probability one for all vectors $\mathbf{a}$. ■

Proof of Lemma 4. This follows from the expansions (34) and (35) and assumptions A3, A7. The magnitudes of S_3 and S_4 follow from Lemmas 1-3 using the arguments employed in Linton (1995, Lemmas 6-9); the moments exist by A2.

■

PROOF OF LEMMA 5. This follows by the same arguments as used in Lemma 4. Note that H_1, H_4, and H_5 are terms that also arise in the expansion for the parametric estimator $\overline{\theta}$, while H_5 is small enough not to contribute to the stated order of magnitude. Finally, by interchanging the order of summation we find that

$$H_2 = \frac{1}{n}\sum_{i=1}^{n} \eta_i^{\dagger}\Omega(X_i)^{-1}D(X_i)' + o_D(n^{-2\epsilon}).$$

■

PROOF OF LEMMA 6. The proof of (36-39) is based on the expansions of Lemmas 4-5 and that under our conditions: $\Pr\left[|\mathcal{Z}_\alpha| > k(\log n)^{1/2}\right] = o(n^{-2\epsilon})$,$\Pr\left[|\mathcal{Z}_{\alpha\beta}| > k(\log n)^{1/2}\right] = o(n^{-2\epsilon})$, $\Pr\left[|\mathcal{Z}_{\alpha\beta\gamma}| > k(\log n)^{1/2}\right] = o(n^{-2\epsilon})$, and $\Pr\left[\sup_{n^{1/2}||\theta-\theta_0||\leq k\log n}|s_{\alpha\beta\gamma\delta}(\theta)| > k(\log n)^{1/2}\right] = o(n^{-2\epsilon})$, for $\alpha,\beta,\gamma,\delta = 1,\ldots,p$, see Bhattacharya and Ghosh (1979, 2.32). We just show the arguments for (39). Write

$$\frac{\partial^3 \widehat{s}_\alpha(\theta)}{\partial\theta_\beta\partial\theta_\gamma\partial\theta_\delta} - \frac{\partial^3 s_\alpha(\theta)}{\partial\theta_\beta\partial\theta_\gamma\partial\theta_\delta} = \frac{1}{n}\sum_{\pi=1}^{p}\sum_{i=1}^{n} a_{\pi i}b_{\pi i}(\theta),$$

where $b_{\pi i}(\theta) = \partial\rho_\pi(Z_i,\theta)/\partial\theta_\beta\partial\theta_\gamma\partial\theta_\delta$ and $a_{\pi i}$ is the π^{th} element of the vector $\widetilde{D}(X_i;\widetilde{\theta})\widetilde{\Omega}(X_i)^{-1} - D(X_i)\Omega(X_i)^{-1}$ [$a_{\pi i}$ depends on the nonparametric estimates, but not on θ]. Further, write $a_{\pi i} = \sum_{\ell=1}^{4} a_{\ell\pi i}$, where $a_{1\pi i}$ is the corresponding element of $\{\widetilde{D}(X_i;\theta_0) - D(X_i)\}\Omega(X_i)^{-1}$, $a_{2\pi i}$ is the corresponding element of $\{\widetilde{D}(X_i;\widetilde{\theta}) - \widetilde{D}(X_i;\theta_0)\}\Omega(X_i)^{-1}$, $a_{3\pi i}$ is the corresponding element of $D(X_i)(\Omega(X_i)^{-1} - \widetilde{\Omega}(X_i)^{-1})$, and $a_{4\pi i}$ is the corresponding element of $\{\widetilde{D}(X_i;\widetilde{\theta}) - D(X_i)\}(\Omega(X_i)^{-1} - \widetilde{\Omega}(X_i)^{-1})$. Then, using the Cauchy-Schwarz inequality we have

$$\sup_{n^{1/2}|\theta-\theta_0|\leq k\log n}\left|\frac{\partial^3\widehat{s}_\alpha}{\partial\theta_\beta\partial\theta_\gamma\partial\theta_\delta}(\theta) - \frac{\partial^3 s_\alpha}{\partial\theta_\beta\partial\theta_\gamma\partial\theta_\delta}(\theta)\right| \leq \left\{\frac{1}{n}\sum_{i=1}^{n} t_i\right\}^{1/2}\sum_{\pi=1}^{p}\sum_{\ell=1}^{4}\left\{\frac{1}{n}\sum_{i=1}^{n} a_{\ell\pi i}^2\right\}^{1/2} \tag{63}$$

where $t_i = \max_\pi \sup_{n^{1/2}|\theta-\theta_0|\leq k\log n}|b_{\pi i}(\theta)|^2$ are independent across i and have $2+\delta$ moments. Therefore, we can apply standard moderate deviation results to $\frac{1}{\sqrt{n}}\sum_{i=1}^{n}\{t_i - E(t_i)\}$, such as in Michel (1974), to conclude that

$$\frac{1}{n}\sum_{i=1}^{n} t_i = \frac{1}{n}\sum_{i=1}^{n} E(t_i) + o_D(n^{-(1/2-\eta)}, n^{-2\epsilon}), \tag{64}$$

i.e., the term $\{\frac{1}{n}\sum_{i=1}^{n} t_i\}^{1/2}$ in (63) can be treated as a constant. Furthermore, we have

$$\{\frac{1}{n}\sum_{i=1}^{n} a_{1\pi i}^2\}^{1/2} \leq k\left[\max_{\alpha,\beta}\max_{1\leq i\leq n}|(V_{ni})_{\alpha\beta}| + \max_{\alpha,\beta}\max_{1\leq i\leq n}|(B_{ni})_{\alpha\beta}|\right] = o_D(n^{-(\epsilon-\eta)}, n^{-2\epsilon}) \tag{65}$$

by Lemma 2. Combining (64) and (65), we get that (39) is true. The term $\{\frac{1}{n}\sum_{i=1}^{n} a_{3\pi i}^2\}^{1/2}$ can be shown to be $o_D(n^{-(1/2-\eta)}, n^{-2\epsilon})$ using standard techniques, while $\{\frac{1}{n}\sum_{i=1}^{n} a_{4\pi i}^2\}^{1/2}$ is of even smaller order.

■

REFERENCES

Abramovitch, L., and K. Singh (1985): "Edgeworth corrected pivotal statistics and the bootstrap," Annals of Statistics 13, 116-132.

Amemiya, T., (1974): "The non-linear two-stage least squares estimator," Journal of Econometrics 2, 105-110.

Andrews, D.W.K., (1991): "Heteroskedasticity and Autocorrelation Consistent Covariance Matrix Estimation." Econometrica 59, 817-858.

Barndorf-Nielsen, O.E. (1989). Asymptotic tecniques for use in statistics. Chapman and Hall: London.

Bather, J. (1997): "A conversation with Herman Chernoff," Statistical Science 11, 335-350.

Besicovitch, A.S. (1993): "On the sum of digits of real numbers represented in the dyadic system," in Classics in Fractals Ed. G.A. Elgar, Addison-Wesley: Reading, MA.

Bhattacharya, R. N. and J. K. Ghosh (1978): "On the Validity of the Formal Edgeworth Expansion," Annals of Statistics 6, 434–451.

Bhattacharya, R. N. and Ranga Rao (1976): Normal approximation and asymptotic expansions. Wiley, New York.

Bickel, P.J. (1982): "On adaptive estimation," Annals of Statistics 10, 647-671.

Bickel, P. J., F. Götze and W. R. Van Zwet (1986): "The Edgeworth Expansion for U Statistics of Degree Two," Annals of Statistics 14, 1463–1484.

Brown, B.W. and W.K. Newey (1996): "Bootstrapping for GMM," Preprint, Rice University.

Callaert, H., P. Janssen and N. Veraverbeke (1980): "An Edgeworth Expansion for U-Statistics," Annals of Statistics 8, 299–312.

Carroll, R.J., and W. Härdle, (1989): "Second Order Effects in Semiparametric Weighted Least Squares Regression." Statistics, 2, 179-186.

Chamberlain, G. (1987): "Asymptotic efficiency in estimation with conditional moment restrictions," Journal of Econometrics 34, 305-334.

Chandra, T.K., and J.K. Ghosh (1979): "Valid asymptotic expansions for the likelihood ratio statistic and other perturbed chi-squared variables," Sankhya, Series A 41, 22-47.

Davies, J.A. (1968): "Convergence rates for probabilities of moderate deviations," The Annals of Mathematical Statistics 39, 2016-2028.

De Jong, P. (1987): "A central limit theorem for generalized quadratic forms," Probability Theory and Related Fields 75, 261-277.

Fan, J. (1992): "Design-Adaptive Nonparametric Regression." Journal of the American Statistical Association, 87, 998-1004.

Fan, J. (1993): "Local Linear Regression Smoothers and their Minimax Efficiencies," The Annals of Statistics, 21, 196-216.

Fan, J., and I. Gijbels (1996), Local Polynomial Modelling and Its Applications Chapman and Hall.

Fan, Y., and O.B. Linton (1996): "Consistent model specification tests: Omitted variables and semiparametric functional forms," Econometrica 64, 865-890.

Fan, Y., and O.B. Linton (1997): "Some higher-order theory for a consistent nonparametric model specification test," Cowles Foundation Discussion Paper no. 1148. Forthcoming in Journal of Statistical Planning and Inference.

Götze, F. (1987). "Approximations for Multivariate U-statistics," Journal of Multivariate Analysis 22, 212-229.

Gozalo, P., and O.B. Linton (1995): "Using Parametric Information in Nonparametric Regression," Forthcoming in The Journal of Econometrics.

Härdle, W., J. Hart, J. S. Marron, and A. B. Tsybakov (1992): "Bandwidth Choice for Average Derivative Estimation," Journal of the American Statistical Association, 87, 218-226.

Härdle, W., and A. B. Tsybakov (1993): "How sensitive are Average Derivatives," Journal of Econometrics, 58, 31-48.

Härdle, W., and O.B. Linton (1994): "Applied nonparametric methods," The Handbook of Econometrics, vol. IV, eds. D.F. McFadden and R.F. Engle III. North Holland.

Härdle, W. and E. Marron (1991) "Bootstrap Simultaneous Error Bars for Nonparametric Regression," Annals of Statistics, 19, 778-796.

Hall, P. (1992): The Bootstrap and Edgeworth Expansion. Springer-Verlag: Berlin.

Hall, P., and J. Horowitz (1996): "Bootstrap critical values for tests based on Generalized Methods of Moments estimation," Econometrica 64, 891-916.

Horowitz, J., (1995): "Bootstrap methods in econometrics: Theory and numerical performance," in Advances in Economics and Econometrics: 7th World Congress, D. Kreps and K.W. Wallis, eds., Cambridge: Cambridge University Press, forthcoming.

Horowitz, J., (1996): "Bootstrapping the smoothed maximum score estimator," mimeo, University of Iowa.

Horowitz, J., (1998): "Bootstrapping in median regression," Econometrica.

Hsieh, D.A., and C.F. Manski (1987): "Monte Carlo Evidence on Adaptive Maximum Likelihood Estimation of a Regression." Annals of Statistics, 15, 541-551.

Jennen-Steinmetz, C., and T. Gasser (1988): "A Unifying Approach to Nonparametric Regression Estimation," Journal of the American Statistical Association 83, 1084-1089.

Jones, M.C., J.S. Marron, and S.J. Sheather (1992): "Progress in data-based bandwidth selection for kernel density estimation," University of New South Wales working paper no 92-04.

Liang, H. (1994): "The Berry-Esséen bounds of error variance estimation in semiparametric regression models," Communications in Statistics 23, 3439-3451.

Liang, H. (1995): "On Bahadur asymptotic efficiency of the maximum likelihood estimator for a generalized semiparametric model," Statistica Sinica 5, 363-371.

Liang. H., and P. Cheng (1993): "Second order asymptotic efficiency in a partially linear model," Statistics and Probability Letters (1993) 18, 73-84.

Linton, O.B. (1995): "Second Order Approximation in the Partially Linear Regression Model," Econometrica 63, 1079-1112.

Linton, O.B. (1996a): "Second order approximation in a linear regression with heteroskedasticity of unknown form," Econometric Reviews 15, 1-32.

Linton, O.B. (1996b): "Edgeworth Approximation for MINPIN Estimators in Semiparametric Regression Models." Econometric Theory 12, 30-60.

Linton, O.B. (1997): "Second-Order approximation for semiparametric instrumental variable esitmators and test statistics." Cowles Foundation Discussion Paper no 1151.

Linton, O.B. and Z. Xiao (1997): "Second order approximation in a semiparametric binary choice model." Manuscript, Yale University.

Masry, E. (1996a): "Multivariate local polynomial regression for time series: Uniform strong consistency and rates," J. Time Ser. Anal. 17, 571-599.

Masry, E. (1996b): "Multivariate regression estimation Local polynomial fitting for time series," Stochastic Processes and their Applications 65, 81-101.

Michel, R. (1974): "Results on probabilities of moderate deviations," The Annals of Probability 2, 349-353.

Müller, H.G., (1988). Nonparametric Regression analysis of Longitudinal Data. Springer Verlag: Berlin.

Nagar, A.L. (1959): "The bias and moment matrix of the general k-class estimator of the parameters in simultaneous equations," Econometrica 27, 575-595.

Newey, W.K., (1986): "Efficient estimation of models with conditional moment restrictions," mimeo, Princeton University.

Newey, W.K., (1988): "Adaptive estimation of regression models via moment restrictions, Journal of Econometrics.

Newey, W.K., (1990): "Efficient instrumental variables estimation of nonlinear models," Econometrica 58, 809-837.

Newey, W.K., (1996): "Optimal choice of instruments in nonlinear models," mimeo.

Nishiyama, Y., and Robinson, P. M. (1997): "Edgeworth expansions for semiparametric averaged derivatives," Forthcoming in Econometrica .

Pfanzagl, J., (1980): "Asymptotic Expansions in Parametric Statistical Theory," in: Developments in Statistics, vol3. ed. P.R.Krishnaiah. Academic Press.

Phillips, P.C.B. and J.Y. Park (1988): "On the formulation of Wald tests of nonlinear restrictions," Econometrica 56, 1065-1083.

Powell, J.L., and T.M. Stoker (1996): "Optimal bandwidth choice for density-weighted averages," Journal of Econometrics 75, 291-316.

Rilestone, P., V.K. Srivastava, and A. Ullah (1996): "The second-order bias and mean squared error of nonlinear estimators," Journal of Econometrics 75, 369-395.

Robinson, P. M. (1987): "Asymptotically Efficient Estimation in the Presence of Heteroscedasticity of Unknown form." Econometrica 56, 875-891.

Robinson, P. M. (1988a): "The Stochastic Difference between Econometric Statistics," Econometrica, 56, 531-548.

Robinson, P. M. (1988b): "Root-N-Consistent Semiparametric Regression," Econometrica, 56, 931-954.

Robinson, P. M. (1991a): "Automatic Frequency Domain Inference on Semiparametric and Nonparametric Models." Econometrica, 59, 1329-1364.

Robinson, P. M. (1991b): "Best Nonlinear Three-stage Least Squares Estimation of certain Econometric Models." Econometrica, 59, 755-786.

Robinson, P. M. (1995): "The normal approximation for semiparametric averaged derivatives," Econometrica 63, 667-680.

Rothenberg, T., (1984): "Approximating the Distributions of Econometric Estimators and Test Statistics." Ch.14 in: Handbook of Econometrics, vol 2, ed. Z. Griliches and M.Intriligator. North Holland.

Rubin, H. and J. Sethuraman (1965): "Probabilities of moderate deviation," Sankhyæ, Series A 27, 325-346.

Silverman, B. (1986): Density estimation for statistics and data analysis. London, Chapman and Hall.

Xiao, Z. and O.B. Linton (1997): "Second order approximation for an adaptive estimator in a linear regression." Forthcoming in Econometric Theory.

Xiao, Z. and P.C.B. Phillips (1996): "Higher order approximation for a frequency domain regression estimator," Journal of Econometrics, 86, 297-336.

D Figure Information

In each figure we give the rejection frequency as a function of bandwidth. The lines give the rejection frequency of the test computed on a grid of bandwidths [and interpolated], while the symbols give the rejection frequency of the automatic method [the horizontal location of the symbols is just determined to make the graphs easy to read.

The solid line represents kernel and the dotted line is for the local linear. The higher line is always the $n = 100$ case and the lower one therefore the $n = 200$ one.

The triangle symbol represents the logit rule-of-thumb, square represents the quadratic probability model rule-of-thumb, while circle represents the nonparametric plug-in. Solid symbols represent kernel and hollow symbols are local linear. The left-most symbols on each graph are for the $n = 100$ and the right-most symbols are for $n = 200$.

Figure 1 is the case where $c = 1$ and hypothesis (a), Figure 2 is the case $c = 0$ with the same null hypothesis, Figure 3 is the hypothesis (b) with $c = 1$. The letter A corresponds to 10% nominal level, B is the 5% case, and C is the 1% case. We have shown Figures 1A,1B,1C, 2A, and 3A.

Figure 1A

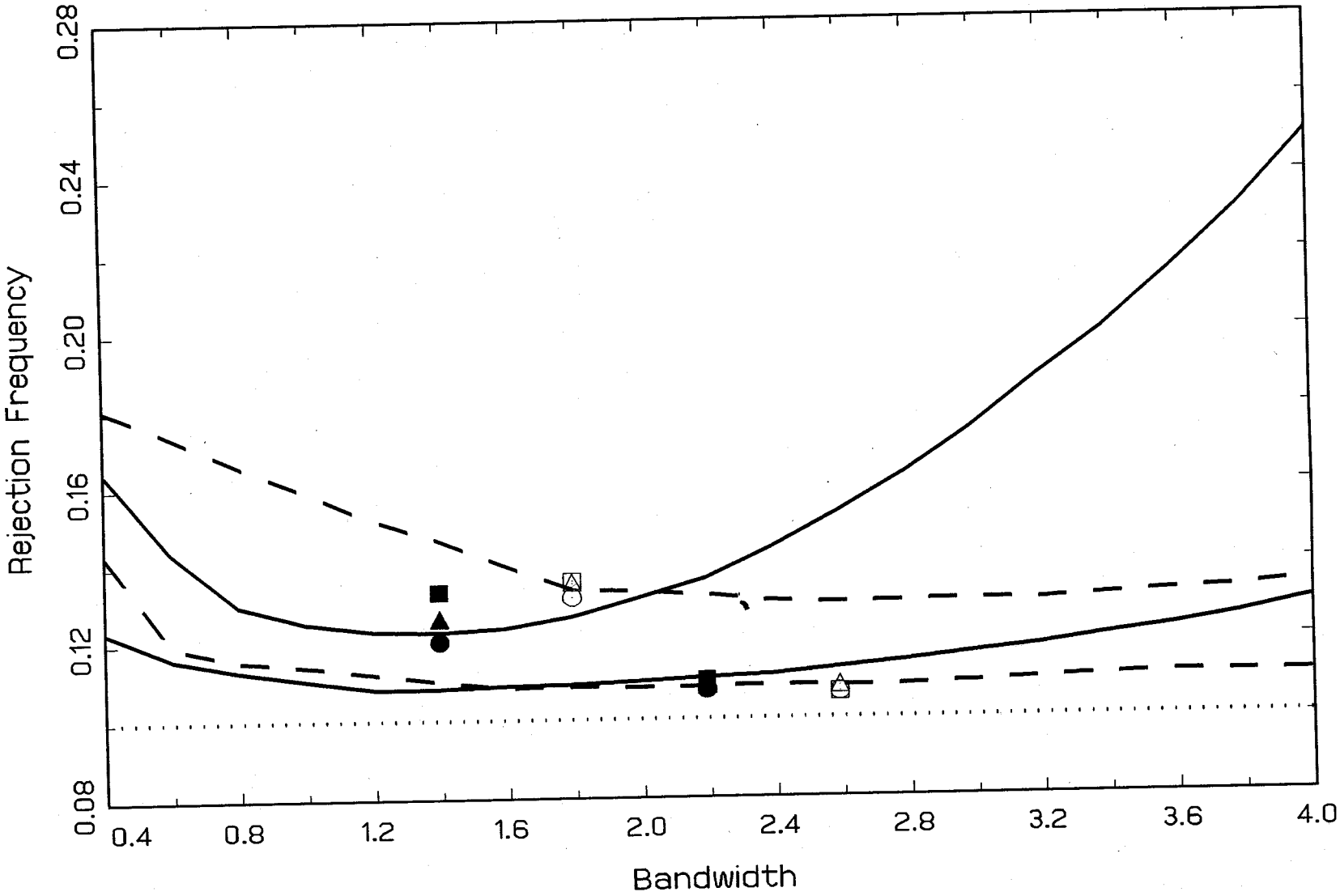

Rejection Frequency
0.28
0.24
0.20
0.16
0.12
0.08
0.4
0.8
1.2
1.6
2.0
2.4
2.8
3.2
3.6
4.0
Bandwidth

Figure 1B

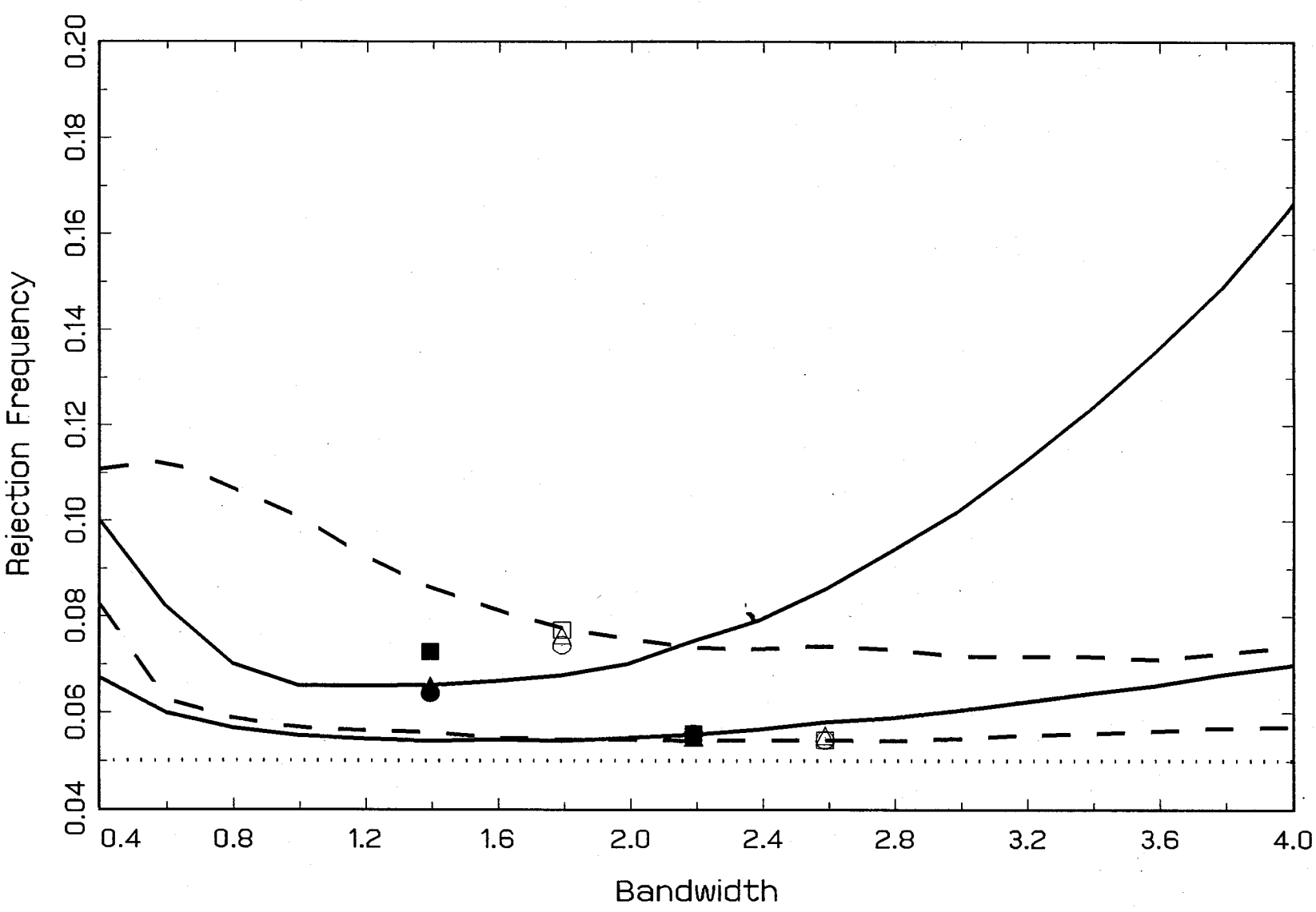

Rejection Frequency
0.04
0.06
0.08
0.10
0.12
0.14
0.16
0.18
0.20
0.4
0.8
1.2
1.6
2.0
2.4
2.8
3.2
3.6
4.0
Bandwidth

Figure 1C

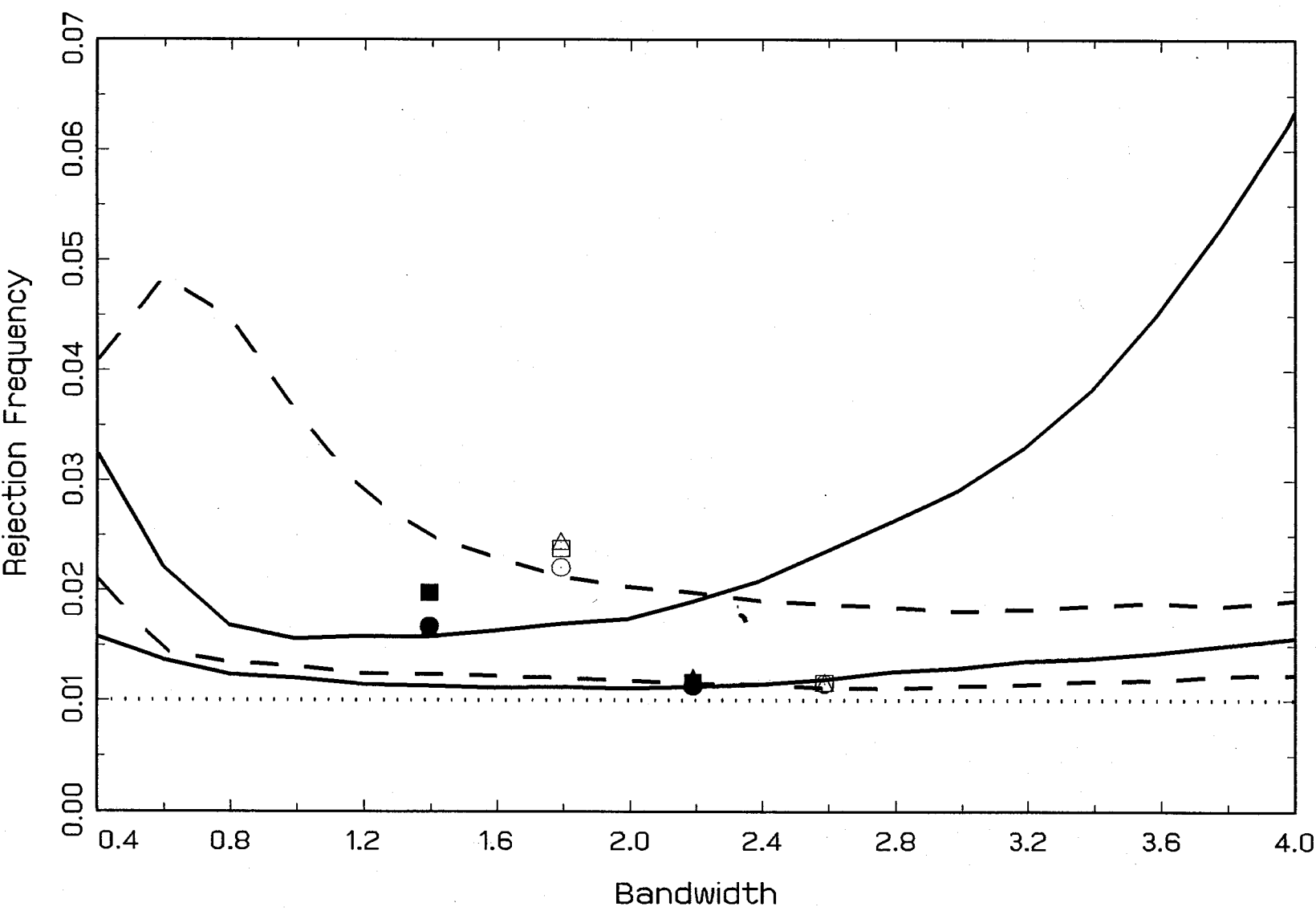

Rejection Frequency
0.00
0.01
0.02
0.03
0.04
0.05
0.06
0.07
0.4
0.8
1.2
1.6
2.0
2.4
2.8
3.2
3.6
4.0
Bandwidth

Figure 2A

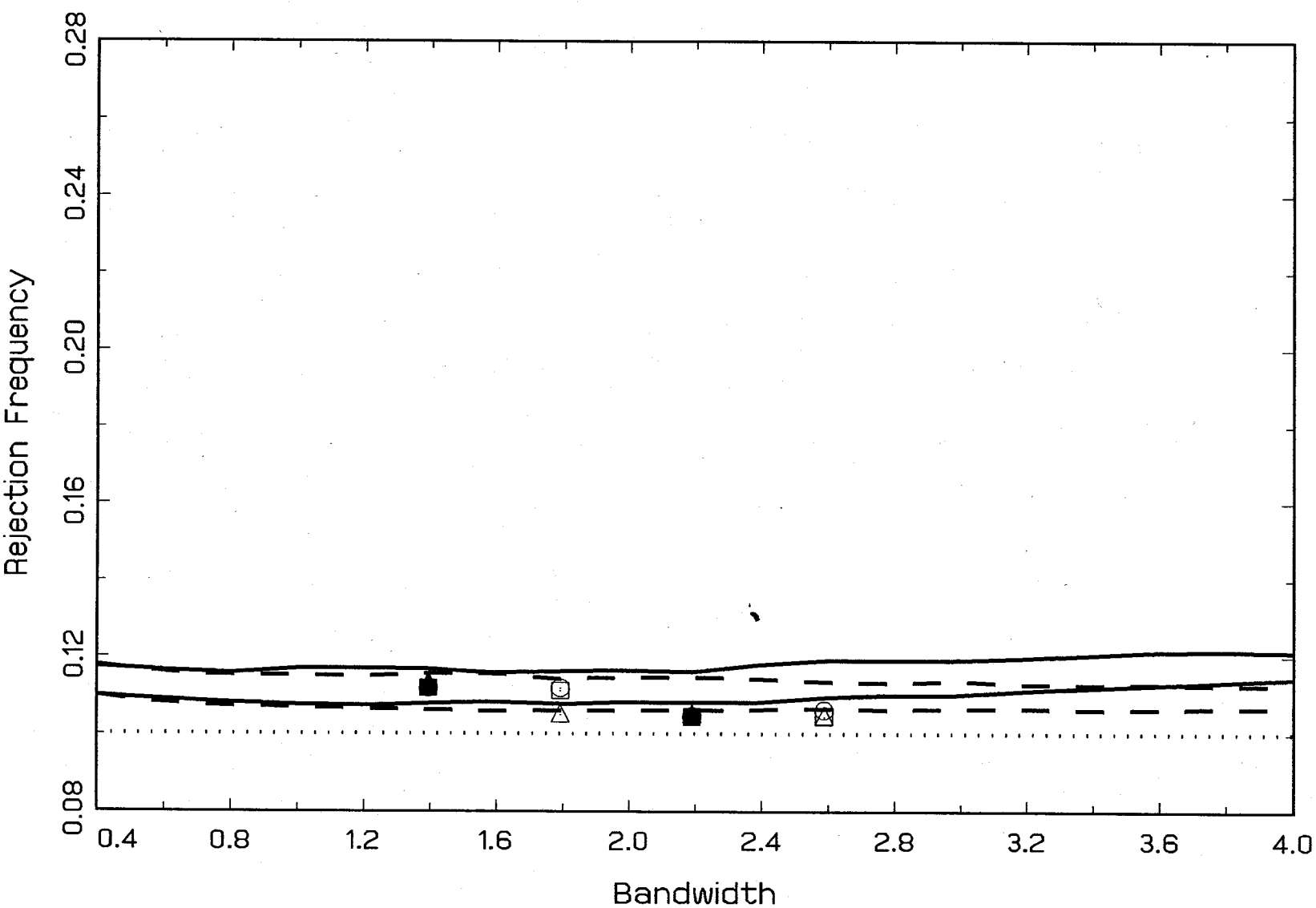
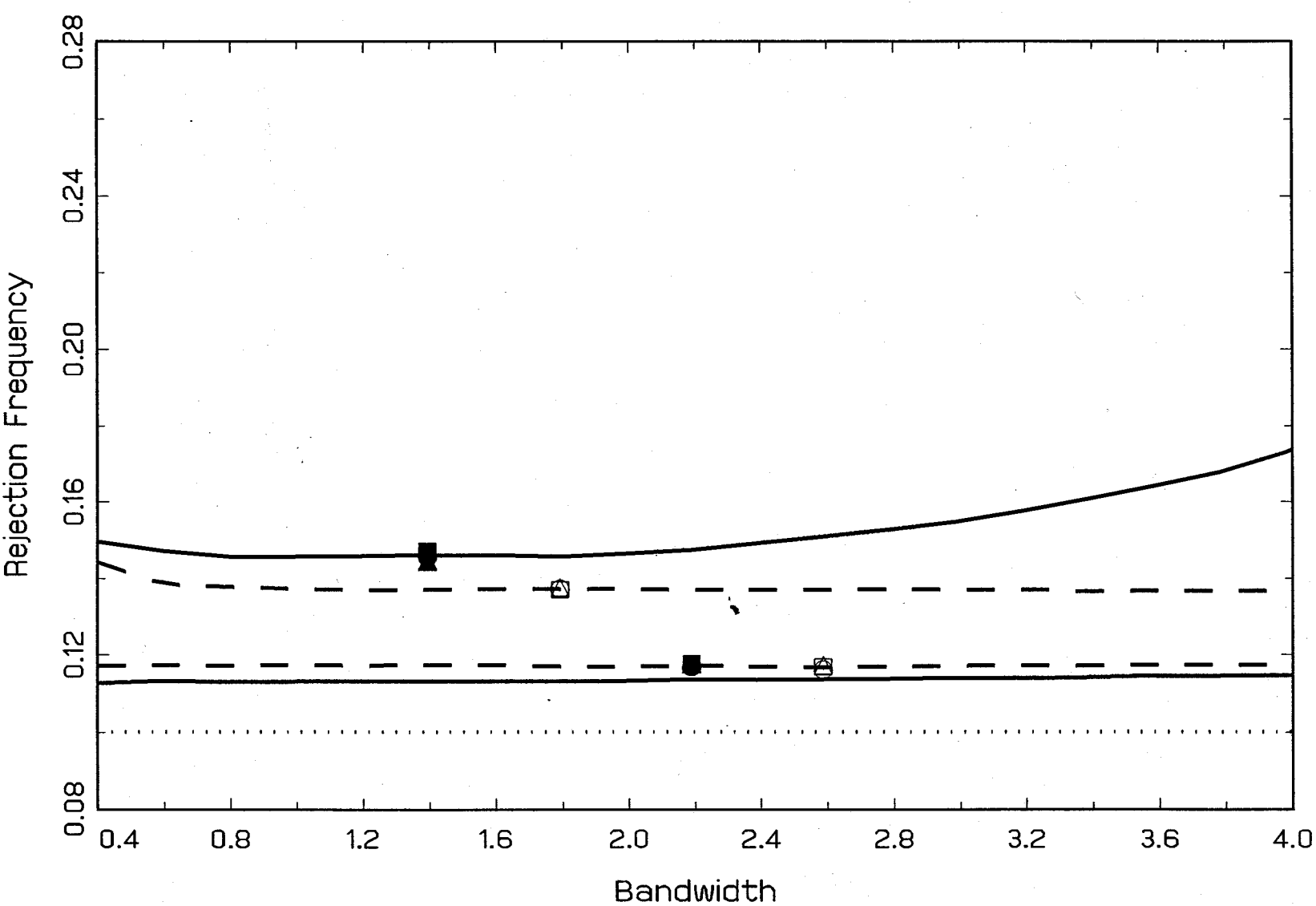

Rejection Frequency
0.28
0.24
0.20
0.16
0.12
0.08
0.4
0.8
1.2
1.6
2.0
2.4
2.8
3.2
3.6
4.0
Bandwidth

Figure 3A


Rejection Frequency
0.28
0.24
0.20
0.16
0.12
0.08
0.4
0.8
1.2
1.6
2.0
2.4
2.8
3.2
3.6
4.0
Bandwidth