Bias-adjusted Nearest Neighbor Estimation for the Partial Linear Model[*]

Guido Imbens
UC Berkeley and NBER

Jack Porter
University of Wisconsin - Madison

First draft: December 2002
This draft March 2005

Abstract

In semiparametric models estimation methods interest is often in the finite dimensional parameter, with the nonparametric component a nuisance function. In many examples, including Robinson's partial linear model and the estimation of average treatment effects, the nuisance function is a conditional expectation. For the large sample properties of the estimators of the parameters of interest it is typically important that the estimators for these nuisance functions satisfy certain bias and variance properties. Estimators that have been used in these settings include series estimators and higher order kernel methods. In both cases the smoothing parameters have to be choosen in a sample-size dependent manner. On the other hand, nearest neighbor methods with a fixed number of neighbours do not rely on sample size dependent smoothing parameters, but they often violate the conditions on the rate of the bias unless the covariates in the regression are of very low dimension. In many cases only scalar covariates are allowed. In this paper we develop an alternative method for estimating the unknown regression functions that, like nearest neighbor methods, does not rely on sample-size dependent smoothing parameters, but that, like the series and higher order kernel methods, does not suffer from bias-rate problems. We do so by combining nearest neighbor methods with local polynomial regression using a fixed number of neighbors.

JEL Classification: .

Keywords: *Matching, Partially Linear Model, Nonparametric Estimation, Nearest Neighbor Estimation*

[*] We are grateful to Jim Powell for helpful discussions. Financial support for this research was generously provided through NSF grant SES 0136789 (Imbens). Electronic correspondence: imbens@econ.berkeley.edu, http://elsa.berkeley.edu/users/imbens/, and jporter@harvard.edu

1 Introduction

In semiparametric models estimation methods interest is often in the finite dimensional parameter, with the nonparametric component a nuisance function. The parameter of interest can often be estimated at parametric (that is $\sqrt{N}$) rates even if the nonparametric component cannot (Newey, 1994). In many examples, including Robinson's partial linear model (Robinson, 1988), and the estimation of average treatment effects (e.g., Rosenbaum and Rubin, 1983; Hahn, 1998; Heckman, Ichimura and Todd, 1998; Hirano, Imbens and Ridder, 2001), the nuisance function is a conditional expectation. For the large sample properties of the estimators of the parameters of interest it is typically important that the estimators for these nuisance functions satisfy certain bias and variance properties. Methods that can generally be constructed to satisfy these conditions include series estimators and higher order kernel methods. In order for these methods to have the desired large sample properties smoothing parameters have to be choosen in a sample-size dependent manner. This requirement has slowed the applicability of these methods as few results are available governing the choice of such smoothing parameters. On the other hand, nearest neigbhour methods with a fixed number of neighbours do not rely on sample size dependent smoothing parameters, but they often violate the conditions on the rate of the bias when the covariates in the regression are of high enough dimension (e.g., Honoré and Estes; Yatchew, 1999; Abadie and Imbens, 2004). In this paper we develop an alternative method for estimating the unknown regression functions that, like nearest neighbor methods, does not rely on sample-size dependent smoothing parameters, but that, like the series and higher order kernel methods, does not suffer from bias-rate problems. We do so by combining nearest neighbour methods with local polynomial regression. We show that with a fixed number of neighbors local polynomial regression can control both the bias and the variance provided the nearest neighbour points are choosen carefully so as satisfy specific conditions on their distribution. In the cases we consider this leads to $\sqrt{N}$ consistent estimators for the parameters of interest with zero asymptotic bias.

The first example of the setting we have in mind is the partial linear model and average treatment effects. Robinson (1988) considers the model:

$$\mathbb{E}[Y|X, Z] = X'\beta + g(Z).$$

Robinson suggests estimating the parameter of interest, β, by first estimating nonparamtrically the two conditional expectations $\mu_Y(Z) = \mathbb{E}[Y|Z]$ and $\mu_X(Z) = \mathbb{E}[X|Z]$, followed by estimating β by least squares regression of the residuals from those conditional expectations:

$$\hat{\beta} = \left(\sum_{i=1}^{N} (X_i - \hat{\mu}_X(Z_i))(X_i - \hat{\mu}_X(Z_i)) \right)^{-1} \left(\sum_{i=1}^{N} (X_i - \hat{\mu}_X(Z_i))(Y_i - \hat{\mu}_Y(Z_i)) \right).$$

The two conditional expectations $\mu_Y(z)$ and $\mu_X(z)$ can be estimated using series or kernel methods. If kernel methods are used higher order kernels are required to ensure that that the square of the bias gets dominated by the variance in the asymptotic distribution of $\sqrt{N}(\hat{\beta} - \beta)$.

In both these examples one can obtain root-N consistent estimators for the parameters of interest, β in the partial linear model case and τ in the average treatment effect case. In order to obtain a root-N consistent estimator one needs to estimate the unknown regression functions. For the properties of the subsequent estimators for the finite dimensional parameters of interest it is important that these nonparametric estimators have particular bias and variance properties. Specifically, the bias needs to be of sufficiently low order so that it gets dominated by the variance in the asymptotic distribution of the parameters of interest. As we shall see, the variance of the estimators of the regression function need not shrink to zero for the estimator of the parameter of interest to be consistent, but it is important that it does not increase too fast with the sample size.

We can distinguish two approaches to the estimation of the nuisance function. First, one can consistently estimate the regression function at each point. The estimators proposed by Robinson (1988), Hahn (1998), and Heckman, Ichimura and Todd (1998) fall into this category. Such estimators can be obtained in a variety of ways, for example through series estimation or kernel methods (or local polynomial versions of this). If one uses kernel methods, one typically needs to use higher order kernels to control the rate at which the bias goes to zero. With series approaches one needs to increase the number of terms in the series sufficiently fast to eliminate the bias at a high enough rate. A complication with these methods is that one needs to choose a sample size dependent smoothing parameter (the bandwidth for the kernel estimators and the number of terms in the series approaches). This can be difficult in practice, with few practical guidelines towards optimal choice of bandwidth. Most of the rules of thumb that have been suggested have optimality properties for estimating the nonparametric function which does not necessarily lead to attractive estimators for the finite dimensional parameter of interest. Exceptions are the papers by Ichimura and Linton (2004) and Imbens, Newey and Ridder (2004) who suggest choices for smoothing parameters specific for average treatment effect estimators.

The second approach is to use estimators with non-vanishing variances. In both the partial linear model and in the average treatment effect case the parameters of interest are estimated by averaging the estimated regression functions, it is not essential that the latter be estimated consistently, merely that the particular average is estimated consistently. Estimators of this type include nearest-neighbor estimators with a fixed number of neighbours. The leading example is the single nearest neighbour estimator where $\mu(X_j) = \mathbb{E}[Y|X_i]$ is estimated as $Y_{\ell(i)}$, where $\ell(i)$ is the index of the closest neighbour for unit j:

$$\ell(i) = \arg \min_{j=1,\ldots,N, j\neq i} \|X_j - X_i\|.$$

As the sample size increases its bias will go to zero and its variance to the conditional variance of Y given X_j, $V(Y|X_i)$. Clearly $Y_{\ell(i)}$ is not consistent for $\mu(X_i)$. Nevertheless using the nearest neighbor approach leads to a consistent estimator for the average treatment effect (see Abadie and Imbens, 2002). The averaging over the estimated regression functions ensures the consistency of $\hat{\tau}$ even if $\hat{\mu}_w(x)$ is not consistent. An attraction of this method is that one does not need to choose a bandwidth that changes with the sample size. (Although the number of nearest neighbors can be viewed as a choice of smoothing parameters, it is a choice that need not change with the sample size.) A problem with this nearest neighbor or matching approach is that with high dimensional covariates the bias in the estimates of the regression functions can be too large relative to the variance. Abadie and Imbens (2002) show that in general the bias term is $O_p(N^{-1/K})$ where K is the dimension of the covariates. Hence with more than one covariates for any root-N consistent estimator the bias will not get dominated, and in fact with more than two continuous covariates the bias will dominate the variance. Abadie and Imbens (2002) combine the matching with a nonparametric bias adjustment based on a consistent estimator of the regression function.

In this paper we propose a bias adjustment to the regression function that maintains the bandwidth free spirit of the nearest neighbor estimator while at the same time controlling the bias up to the desired degree. It therefore combines some of the attractive features of kernel and series estimators in terms of asymptotic properties with those of nearest neighbor estimators in terms of ease of implementation. Our proposed estimator is based on selecting a fixed number of neigbors and using those to reduce the bias by least squares methods. A naive way of implementing this by choosing directly the nearest neighbors is shown not to be valid because the variance of such an estimator explodes. Our proposed estimator controls the variance behavior by modifying the selection of the neighbors through requiring them to be spread around the point where we wish to approximate the regression function.

2 The Scalar Covariate Case

To fix ideas and provide some intuition for the main contributions of the paper, we discuss in this section a special case with a scalar covariate. The substantive results in this section are of less direct interest because in the scalar covariate case the nearest neighbor method has a sufficiently small bias that it does not enter into the asymptotic distribution in both the partial linear model and in the average treatment examples. Nevertheless, the approache to lowering the stochastic order of the bias still applies, and the methods discussed here provide the key to the general case.

Consider first the problem of estimating a regression function $\mu(x) = \mathbb{E}[Y|X = x]$ at a particular point, say at $x = 0$ so that we are interested in $\mu(0)$. Let $\varepsilon_i = Y_i - \mu(X_i)$, with $\varepsilon_i \perp X_i$ and $V(\varepsilon_i|X_i) = \sigma^2$. Assume throughout that the density of X can be bounded away from zero in a neighborhood of zero. Suppose we have an i.i.d. sample $(Y_1, X_1), \ldots, (Y_N, X_N)$. Let

$$\ell_1 = \arg\min_{i=1,\ldots,N} |X_j|,$$

and more generally let ℓ_m be the index of the observation that is the m^{th} closest to 0 in terms of the covariate.

First consider a simple nearest neighbor estimator for $\mu(0)$ based on a single neighbor. Begin by finding the observation with the value of X closest to zero, that is, unit ℓ_1 with covariate value X_{ℓ_1}. Then the first estimator for $\mu(0)$ is

$$\hat{\mu}(0) = Y_{\ell_1}.$$

The properties of this estimator follow from the results in Abadie and Imbens (2004). As the sample size gets large,

Next, consider a nearest neighbor estimator for $\mu(0)$ based on the two nearest neighbors. Begin by finding the two observations with values of X closest to zero, that is, units ℓ_1 and ℓ_2, with covariate values X_{ℓ_1} and X_{ℓ_2}. The estimator is defined by averaging the corresponding dependent variable observations,

$$\hat{\mu}(0) = \frac{1}{2} \cdot Y_{\ell_1} + \frac{1}{2} \cdot Y_{\ell_2}.$$

By averaging only two observations, such an estimator is necessarily inconsistent. To illustrate the key issues with our proposed matching estimator, we focus on the properties of the conditional bias and variance.

$$\begin{aligned} \hat{\mu}(0) - \mu(0) &= \frac{1}{2} \cdot Y_{\ell_1} + \frac{1}{2} \cdot Y_{\ell_2} - \mu(0) \\ &= \left(\frac{1}{2} \cdot [\mu(X_{\ell_1}) + \mu(X_{\ell_2})] - \mu(0) \right) && (2.1) \\ &\quad + \frac{1}{2} \cdot (\varepsilon_{\ell_1} + \varepsilon_{\ell_2}) && (2.2) \end{aligned}$$

The leading term (2.1) determines the bias, while the second term (2.2) determines the conditional variance of the estimator. Assuming smoothness in $\mu(\cdot)$ we can linearize the first term to give the order of bias, as follows

$$\begin{aligned} &\mathbb{E}[\hat{\mu}(0) - \mu(0) | X_1, \ldots, X_N] \\ &\quad= \frac{1}{2} \cdot [\mu(X_{\ell_1}) + \mu(X_{\ell_2})] - \mu(0) \\ &\quad= \frac{1}{2} \cdot \left[\left(\mu(0) + \mu'(0) \cdot X_{\ell_1} + o_p(X_{\ell_1}) \right) + \left(\mu(0) + \mu'(0) \cdot X_{\ell_2} + o_p(X_{\ell_2}) \right) \right] - \mu(0) \\ &\quad= \mu'(0) \cdot \frac{X_{\ell_1} + X_{\ell_2}}{2} + o_p(X_{\ell_1}) + o_p(X_{\ell_2}). && (2.3) \end{aligned}$$

From (2.1), the bias arises from the matching discrepancy between X_{ℓ_1} and X_{ℓ_2} on the one hand and 0 on the other hand. For any finite, fixed m, the matching discrepancy X_{ℓ_m} is $O_p(N^{-1})$. The linearization (2.3) shows more specifically that the bias in $\hat{\mu}(0)$ is $O_p(X_{\ell_1} + X_{\ell_2})$ $= O_p(N^{-1})$.

Under the homoskedasticity assumption maintained in this section the conditional variance is given by

$$V(\hat{\mu}(0) - \mu(0) | X_1, \ldots, X_N) = \frac{\sigma^2}{2}.$$

Using only the two nearest matches the variance is fixed, not converging to zero, implying inconsistency of this estimator.

Now consider a modification of the above estimator. Instead of simply averaging the two nearest neighbors, consider a weighted average of the two nearest neighbors:

$$\begin{aligned}
&\tilde{\mu}(0) - \mu(0) = \rho_1 Y_{\ell_1} + \rho_2 Y_{\ell_2} - \mu(0) \\
&\quad = \quad [\rho_1 \mu(X_{\ell_1}) + \rho_2 \mu(X_{\ell_2})] - \mu(0) \ + \ \rho_1 \varepsilon_{\ell_1} + \rho_2 \varepsilon_{\ell_2} \\
&\quad = \quad \rho_1[\mu(0) + \mu'(0)X_{\ell_1} + \frac{1}{2}\mu''(0)X_{\ell_1}^2 + o_p(X_{\ell_1}^2)] + \rho_2[\mu(0) + \mu'(0)X_{\ell_2} + \frac{1}{2}\mu''(0)X_{\ell_2}^2 + o_p(X_{\ell_2}^2)] - \mu(0) \\
&\quad \qquad + \rho_1 \varepsilon_{\ell_1} + \rho_2 \varepsilon_{\ell_2} \\
&\quad = \quad [(\rho_1 + \rho_2) - 1]\mu(0) + \mu'(0)(\rho_1 X_{\ell_1} + \rho_2 X_{\ell_2}) + \frac{1}{2}\mu''(0)(\rho_1 X_{\ell_1}^2 + \rho_2 X_{\ell_2}^2) + o_p(\rho_1 X_{\ell_1}^2 + \rho_2 X_{\ell_2}^2) \\
&\quad \qquad + \rho_1 \varepsilon_{\ell_1} + \rho_2 \varepsilon_{\ell_2}
\end{aligned}$$

With the aim of reducing bias, we choose the weights to satisfy $\rho_1 + \rho_2 = 1$ and $\rho_1 X_{\ell_1} + \rho_2 X_{\ell_2} = 0$, so that

$$\text{and } \rho_2 = -X_{\ell_1}/(X_{\ell_2} - X_{\ell_1}).$$

Then the conditional bias is

$$E[\tilde{\mu}(0) - \mu(0)|X_1, \ldots, X_N] = \frac{\mu''(0)}{2} \cdot (\rho_1 X_{\ell_1}^2 + \rho_2 X_{\ell_2}^2) + o_p(\rho_1 X_{\ell_1}^2 + \rho_2 X_{\ell_2}^2).$$

This choice of weights eliminates the linear bias term in the bias expansion. However, even though X_{ℓ_1} and X_{ℓ_2} are both of stochastic order $O_p(N^{-2})$, we cannot immediately conclude that the bias is $O_p(N^{-2})$. Such a result depends on the behavior of the weights ρ_1 and ρ_2.

The conditional variance of this estimator is

$$V(\tilde{\mu}(0) - \mu(0)|X_1, \ldots, X_N) = (\rho_1^2 + \rho_2^2)\sigma^2.$$

With $\rho_1 = X_{\ell_2}/(X_{\ell_2} - X_{\ell_1})$ and $\rho_2 = -X_{\ell_1}/(X_{\ell_2} - X_{\ell_1})$, if the magnitude of the denominator $|X_{\ell_2} - X_{\ell_1}|$ is much smaller than the magnitude of either numerator $|X_{\ell_1}|$ or $|X_{\ell_2}|$, then the weights cause the bias and variance to be large.

There is a simple solution in this case. Suppose we take the two nearest neighbours that are on opposite sides of zero. With, say, X_{ℓ_1} negative and X_{ℓ_2} positive, $X_{\ell_2} - X_{\ell_1} \geq \max\{|X_{\ell_1}|, |X_{\ell_2}|\}$ so $0 \leq \rho_1, \ \rho_2 \leq 1$. In this case, we can conclude that the bias is indeed $O_p(N^{-2})$ and the variance is bounded by σ^2.

Next we note an alternative interpretation of the weighted average matching estimator, $\tilde{\mu}(0)$. Given the two nearest neighbours, we fit a line through them and use that line to predict the regression function at zero. Based on the two points (X_{ℓ_1}, Y_{ℓ_1}) and (X_{ℓ_2}, Y_{ℓ_2}) the estimated line is

$$\begin{aligned}
\tilde{\mu}(x) &= \hat{\alpha} + x \cdot \hat{\beta} \\
&= Y_{\ell_1}\left(\frac{X_{\ell_2}}{X_{\ell_2} - X_{\ell_1}}\right) - Y_{\ell_2}\left(\frac{X_{\ell_1}}{X_{\ell_2} - X_{\ell_1}}\right) + x \cdot \frac{Y_{\ell_2} - Y_{\ell_1}}{X_{\ell_2} - X_{\ell_1}},
\end{aligned}$$

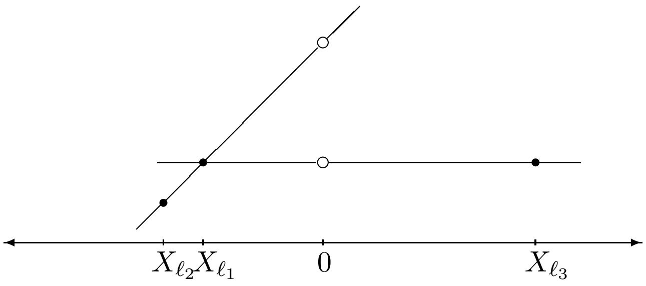

Figure 1: XX

where

$$\hat{\alpha} = Y_{\ell_1}\left(\frac{X_{\ell_2}}{X_{\ell_2} - X_{\ell_1}}\right) - Y_{\ell_2}\left(\frac{X_{\ell_1}}{X_{\ell_2} - X_{\ell_1}}\right) \qquad \text{and} \quad \hat{\beta} = \frac{Y_{\ell_2} - Y_{\ell_1}}{X_{\ell_2} - X_{\ell_1}}.$$

Hence at zero the estimated regression line is

$$\tilde{\mu}(0) = \hat{\alpha} = \rho_1 \cdot Y_{\ell_1} + \rho_2 \cdot Y_{\ell_2},$$

with weights $\rho_1 = X_{\ell_2}/(X_{\ell_2} - X_{\ell_1})$ and $\rho_2 = -X_{\ell_1}/(X_{\ell_2} - X_{\ell_1})$ as before. So the intercept of the regression line through the two nearest neighbors produces a weighted average estimator with weights chosen to eliminate the linear term in the bias expansion. From this derivation it is clear that a line between two close observations on the same side of zero may produce a poor prediction at zero, see Figure 1. One would prefer slightly further away neighbors that are spread out enough from each other to decrease the variance in prediction.

Note that irrespective of how one chooses the points, if the two points do get closer to zero as the sample size increases, then the estimate of the slope coefficient gets large in absolute value with high probability. To see this, write the estimator for the slope coefficient as

$$\hat{\beta} = \frac{Y_{\ell_2} - Y_{\ell_1}}{X_{\ell_2} - X_{\ell_1}} = \frac{\mu(X_{\ell_2}) - \mu(X_{\ell_1})}{X_{\ell_2} - X_{\ell_1}} + \frac{\varepsilon_{\ell_2} - \varepsilon_{\ell_1}}{X_{\ell_2} - X_{\ell_1}}.$$

As the sample size increase, the first term converges to $\mu'(0)$. The second term has expectation zero conditional on $X_1, \ldots, X_N$, but its conditional variance is $\sigma^2/(X_{\ell_2} - X_{\ell_1})^2$ which will get large with probability one. From this one may surmise that adjusting for bias using this estimated regression function need to be very effective. Nevertheless, as we will show more formally, as long as the point at which the regression function is evaluated is not too far from the points at which it is being estimated, this will be an effective adjustment for the bias.

take the discussion all the way to estimator for coefficient in partial linear model and see what the order of the bias for the single match, two averaged matches and linear regression match cases are.

The remainder of the paper consists of formalizing and extending the results discussed above in two ways. First, we allow the covariates to be vectors rather than scalars. Second, we consider bias corrections involving higher order polynomials rather than only linear regressions.

With d-dimensional covariates the simple matching estimator for the regression function at $\mu(0)$ has a bias term of order $O_p(N^{-1/d})$. A local linear regression of the type we discussed above, can lower the bias term to the order $O_p(N^{-})$. This is in general not sufficient to obtain root-N consistent estimators for the parameters of interest in the partial linear model and average treatment effect cases. A local polynomial regression of order p can further reduce the order of the bias to $O_p(N^{-})$. With $p >$, this is sufficient to obtain root-N consistency.

To ensure the appropriate stochastic order for the bias term, one needs to restrict the manner in which the nearest neighbours are selected for use in the estimator. The notion of nearest neigbhours on either side of the point where one is interested in estimating the regression function does not extend naturally to higher dimensions, but one can restrict attention to sets of neighbours that are sufficiently spread out around the point of interest by restricting the implicit weights with these points. As a result we will be able to specify a selection rule for the allowable nearest neighbours in such a way that with a fixed number of neighbours we can obtain estimators for the nuisance function with low enough bias and non-increasing variance, and thus a root-N consistent estimator for the parameter of interest that does not require the choice of a sample-size dependent smoothing parameter.

3 Nearest Neighbor Estimators for Regression Functions with Uniformaly Low Order Bias

3.1 Definitions

Suppose we have a random sample $(X_1, Y_1), (X_2, Y_2), \ldots, (X_N, Y_N)$ of size N. Let $\mathbb{X} \subset \mathbb{R}^K$ be the compact support of X. Define

$$\mu(x) \equiv \mathbb{E}[Y|X = x],$$

and

$$\sigma^2(x) \equiv \mathbb{V}(Y|X = x).$$

In this section we develop an estimator for $\mu(x)$ with a bias of order less than $N^{-1/2}$, and a bounded variance:

$$\sup_{x\in\mathbb{X}} |\mathbb{E}\left[\hat{\mu}(x)|\mathbf{X}\right] - \mu(x)| = o_p\left(N^{-1/2}\right), \tag{3.4}$$

$$\sup_{x\in\mathbb{X}} \mathbb{V}\left(\hat{\mu}(x)|\mathbf{X}\right) = O_p(1). \tag{3.5}$$

The estimator of $\mu(\cdot)$ at x is a least squares predictor based on x and higher order moments, estimated on a fixed number of points. These points are choosen from the observations close to x, with the restriction that they are sufficiently spread out to guarantee that the variance remains bounded with high probability. First we describe the restrictions that the points need to satisfy in order to control the variance. Next we show that there will exist, with probability one in large samples, a sufficient number of points nearby every observed value of x that satisfy these restrictions. Then we show that these points are sufficiently close to remove the bias to a sufficiently high degree.

First we introduce some additional notation. Let $\lambda = (\lambda_1, \ldots, \lambda_K)'$ be a K-dimensional vector of non-negative integers with norm $|\lambda| = \sum_{k=1}^{K} \lambda_j$, and let, for $x \in \mathbb{R}^K$, the function x^λ be equal to $\prod_{k=1}^{K} x_k^{\lambda_k}$. Let Λ_P be the set of λ with $|\lambda| \leq P$. The number of distinct elements in Λ_P is $M(P,K) = (K+P)!/(K!P!)$.[1] Furthermore, let $h_P(x)$ a vector of dimension $M(P,K)$ of functions of K-dimensional vectors x, with each function $h_{P,m}(x)$ equal to a different function x^λ for $\lambda \in \Lambda_P$. So, with $P = K = 2$, $M(P,K) = 6$, and the vector $h_P(x)$ is

$$h_P(x) = \begin{pmatrix} 1 \\ x_1 \\ x_2 \\ x_1^2 \\ x_2^2 \\ x_1 \cdot x_2 \end{pmatrix},$$

or some reordering thereof.

For each nonnegative integer L let $\mathbb{M}_L$ be a subset of $\mathbb{R}^K$ with L elements: $\mathbb{M}_L = \{x_1, \ldots, x_L\}$. Then for a given set $\mathbb{M}_{M(P,K)}$ let $H(\mathbb{M}_{M(P,K)})$ be the $M(P,K) \times M(P,K)$ matrix

$$H\left(\mathbb{M}_{M(P,K)}\right) = \frac{1}{M(P,K)} \sum_{x\in\mathbb{M}_{M(P,K)}} h_P(x) h_P(x)'.$$

Let $\mu_{P,\mathbb{M}_{M(P,K)}}(x)$ be the approximation to $\mu(x)$ based on a Pth order polynomial fitted on the $M(P,K)$ points $(z, \mu(z))$, for $z \in \mathbb{M}_{M(P,K)}$. The approximation to the regression function at x can be written as a weighted average of the value of the function at these $M(P,K)$ points:

$$\mu_{P,\mathbb{M}_{M(P,K)}}(x) = \sum_{z\in\mathbb{M}_{M(P,K)}} \omega_z(x) \cdot \mu(z),$$

with the weight for $z \in \mathbb{M}_{M(P,K)}$ equal to

$$\omega_z(x) = h_P(x)' H(\mathbb{M}_{M(P,K)})^{-1} h_P(z).$$

[1] Choose P out of a set of $K+1$ with replacement without ordering.

If we were to fit the regression function using $M(P,K)$ pairs $\{(X_m, Y_m)\}_{m=1}^{M(P,K)}$, the bias is the same, and the conditional variance would be

$$\mathbb{V} = \sum_{m=1}^{M} \omega_{X_m}^2(x) \cdot \sigma^2(X_m).$$

Hence by bounding the weights, the fact that the conditional variance is bounded will guarantee that the conditional variance of the estimator is bounded.

For a given set $\mathbb{B} \subset \mathbb{R}^K$ and for a set $\mathbb{M}_{M(P,K)}$, let $\omega(\mathbb{B}, \mathbb{M}_{M(P,K)})$ be the maximum weight for predicting $\mu(x)$, with x choosen in the set $\mathbb{B}$ and using the $M(P,K)$ points in $\mathbb{M}$ to approximate $\mu(x)$:

$$\omega(\mathbb{B}, \mathbb{M}_{M(P,K)}) = \max_{z \in \mathbb{M}_{M(P,K)}} \sup_{x \in \mathbb{B}} h_P(x)' H(\mathbb{M}_{M(P,K)})^{-1} h_P(z).$$

The key to ensuring that the variance restriction (3.5) is met is controlling these maximum weights. We shall show that with the volume of $\mathbb{B}$ shrinking to zero as a function of the sample size sufficiently fast to satisfy the bias condition (3.4), there will with probability one be a set $\mathbb{M}_{M(P,K)}$ that is a subset of the intersection of $\mathbb{B}$ and the sample $\{X_1, \ldots, X_N\}$, with all the weights bounded.

In the next subsection we show that such a set $\mathbb{M}_{M(P,K)}$ exists with probability one, and give a bound on the rate at which it is shrinking to zero. In Subsection 3.3 we show that this rate is fast enough to make the bias go to zero faster than $\sqrt{N}$.

3.2 Feasibility

First define the unit ball $\mathbb{B}_K = \{x \in \mathbb{R}^K | \|x\| \leq 1\}$, where $\|x\| = (x'x)^{1/2}$ is the Euclidean distance. Next, consider for each P and K the function $\delta(P,K)$:

$$\delta(P,K) = \inf_{\mathbb{M}_{M(P,K)} \subset \mathbb{B}_K} \omega(\mathbb{B}_K, \mathbb{M}_{M(P,K)}).$$

The function $\delta(P,K)$ is equal to the maximum weight for any point in the unit ball, when the set $\mathbb{M}$ is choosen optimally from elements from the unit ball to minimize this maximum weight. This function $\delta(P,K)$ can be tabulated as a function of the non-negative integer P and the positive integer K. For example, with $P = 0$ and $K = 1$, the maximum weight is $\delta(0,1) = 1$. More generally, with $P = 0$ the function is $\delta(P,K) = 1/K$. For $P = 1$ we have $\delta(1,1) = 1$. There appears to be no general analytic representation for this function.

The following lemma shows that the set $\mathbb{M}$ can always be choosen so that the maximum weight is finite.

Lemma 3.1 (FINITE WEIGHTS)
For all positive integers P and K, $\delta(P,K)$ is finite.

Proof: See Appendix.

Next, define for all $\alpha > 0$ the ball

$$\mathbb{B}_{N,\alpha}(z) = \{x \in \mathbb{X} | \; \|x - z\| \leq N^{-\alpha}\}.$$

Now define for each $x \in \mathbb{X}$, given a sample $X_1, \ldots, X_N$ of size N, and given $\alpha > 0$, the best achievable set of weights within the neighbourhood $\mathbb{B}_{N,\alpha}(x)$:

$$\delta_\alpha(x; X_1, \ldots, X_N) = \min_{\mathbb{M}_{M(P,K)} \subset \left(\{X_1,\ldots,X_N\} \cap \mathbb{B}_{N,\alpha}(x)\right)} \omega(\mathbb{B}_{N,\alpha}(x), \mathbb{M}_{M(P,K)}).$$

If there are fewer than $M(P, K)$ elements in the set $\{X_1, \ldots, X_N\} \cap \mathbb{B}_{N,\alpha}$, the function $\delta_\alpha(x; X_1, \ldots, X_N)$ is defined to be infinite. Again, let us look at the interpretation of this function in more detail. In a shrinking neighbourhood of x, and for a given sample $X_1, \ldots, X_N$ we consider the maximum weight for each point in this neighbourhood given an optimal choice for the set of $M(P, K)$ sample points in this neighbourhood.

Finally, define the set of $x \in \mathbb{X}$ where the weights are arbitrarily close to optimal given the sample.

$$\mathbb{X}^*(\varepsilon, \alpha; X_1, \ldots, X_N) = \{x \in \mathbb{X} | \delta_\alpha(x; X_1, \ldots, X_N) < \delta(P, K) + \varepsilon\}.$$

The next result states that in large samples the weights will satisfy the necessary restrictions for all $x \in \mathbb{X}$.

Theorem 3.1 (Uniformity)
If $\alpha < 1/K$, then for all $\varepsilon, \nu > 0$, there is an $\underline{N}$ such that for $N > \underline{N}$,

$$Pr\Big(\mathbb{X} \subset \mathbb{X}^*(\varepsilon, \alpha; X_1, \ldots, X_N)\Big) > 1 - \nu.$$

Proof: See Appendix.

This theorem states that with probability approaching one as the sample size increases, there will be for each element x of $\mathbb{X}$ a set of points $\mathbb{M}_{M(P,K)} \subset \{X_1, \ldots, X_N\}$ that is increasingly close to x, (all elements of $\mathbb{M}_{M(P,K)}$ are less than $N^{-\alpha}$ away from x), and with maximum weight for the prediction of $\mu(\cdot)$ at x using the $M(P, K)$ elements of $\mathbb{M}_{M(P,K)}$ bounded by $\delta(P, K) + \varepsilon$. The fact that the maximum weight is bounded means that we can control the variance of the prediction.

3.3 Bias Reduction

Consider predicting the regression function $\mu(x)$ at a specific value x, using a Pth order polynomial estimated on the $M(P, K)$ points $\{(x_1, \mu(x_1)), \ldots, (x_{M(P,K)}, \mu(x_{M(P,K)})\}$. Let $\mathbb{M} =$

$\{x_1, \ldots, x_{M(P,K)}\}$ be the set of X-values. Following from the argument given before, the prediction can be written as

$$\hat{\mu}_{P,\mathbb{M}}(x) = \sum_{z\in\mathbb{M}_{M(P,K)}} \omega_z(x)\cdot\mu(z),$$

where

$$\omega_z(x) = h_P(z)'H(\mathbb{M})^{-1}h_P(x).$$

The bias of the prediction is

$$\sum_{z\in\mathbb{M}} \omega_z(x)\cdot(\mu(z)-\mu(x)).$$

Assumption 3.1 (SMOOTHNESS)
The regression function $\mu(x)$ is $P+1$ times continuously differentiable and the absolute value of the $P+1$th derivative of $\mu(x)$ is bounded by C_{P+1}.

Lemma 3.2 (BIAS REDUCTION)
On the set $\mathbb{X}$ the bias $\hat{\mu}_{P,\mathbb{M}}(x)-\mu(x)$ is bounded by

$$\sup_{z\in\mathbb{M}}\sup_{x}|\omega_z(x)|\cdot\sup_{x,z\in\mathbb{X}}\|x-z\|^{P+1}.$$

Proof: See Appendix.

Suppose we have a sample $\{(X_1,Y_1),\ldots,(X_N,Y_N)\}$. Define for a given $\varepsilon>0$,

$$\mathbb{M}(x) = \text{argmin}_{\mathbb{M}_{M(P,K)}\subset\{X_1,\ldots,X_N\}:\omega(\{x\},\mathbb{M}_{M(P,K)})<\delta(P,K)+\varepsilon}\max_{z\in\mathbb{M}_{M(P,K)}}\|x-z\|.$$

Then define the estimated regression function

$$\hat{\mu}_P(x) = \sum_{z\in\mathbb{M}(x)} \omega_z(x)\cdot\mu(z),$$

where $\omega_z(x)$ is as before. In large samples with probability one there will exist for all x a set $\mathbb{M}(x)$ satisfying the conditions in its definition. In finite samples, however, such sets do not necessarily exist. If they do not exist, we will define the estimated regression function to be the value at the single closest neighbour. Since we will only need to use this modification with probability zero, this does not affect any of the results.

Theorem 3.2 (UNIFORM BIAS REDUCTION)
For any $\alpha<1/K$, and any $\eta>0$, there is a $\underline{N}$ such that for all $N>\underline{N}$,

$$Pr\left(\sup_{x\in\mathbb{X}}|\hat{\mu}_P(x)-\mu(x)|<N^{-\alpha(P+1)}\right)>1-\eta.$$

Proof: See Appendix.

Fixing α at some value less than $1/K$, we can always find a P such that the order of the bias is less than $N^{-1/2}$. Specifically, we choose $P > 1/(2\alpha) - 1$. The smallest possible value for P is therefore $P^*(K) = K/2$ if K is even and $P^*(K) = K/2 - 1/2$ if K is odd. Then we are mainly interested in $\delta(P, K)$ at the minimum value for P, $\delta^*(K) = \delta(P^*(K), K)$.

Note also that since the order of the distance to the nearest neighbour is $N^{-1/K}$, one expects to have a finite number of neighbours with neighbourhoods of order $N^{-1/K}$. However, in order to insure that we can find close neighbours for every sample point, we expand the neighbourhood by ε.

4 Bias-adjusted Nearest Neighbor Estimation of Robinson's Partial Linear Model

For each unit i, for $i = 1, \ldots, N$, we observe the triple (X_i, Z_i, Y_i). We are interested in the conditional expectation of Y given X and Z. The model postulates additive separability of this conditional expectation between X and Z, a linear regression function for X:

$$Y_i = X_i'\beta + g(Z_i) + \varepsilon_i,$$

with $\mathbb{E}[\varepsilon_i | X_i, Z_i] = 0$. In order to estimate β, we wish to regress Y on X, both in deviations from the estimated conditional expectation given Z.

Let K be the dimension of Z, and L the dimension of X. The polynomial adjustment we use is of order $P = P^*(K) = [(K-1)/2]$, where $[A]$ is the largest integer less than or equal to A. Let $M^* = M(P, K) = M(P^*(K), K)$ be the required number of nearest neighbors. Let $\delta^* = \delta(P, K) = \delta(P^*(K), K)$ be the optimal feasible weight. We fix $\delta > \delta^*$ to be the required weight. Then for each i we look for a set $\mathcal{M}$ of M^* neighbors such that

$$\max_{j \in \mathcal{M}} h_P(z_j) \left(\sum_{l \in \mathcal{M}} h_P(z_l) h_P(z_l)' \right)^{-1} \sum_{l \in \mathcal{M}} h_P(z_l) \leq \delta,$$

Among the sets satisfying this condition we look for the one that minimizes

$$\sum_{j \in \mathcal{M}} \| Z_j - Z_i \|.$$

Given these sets $\mathcal{M}(i)$ we estimate $\mathbb{E}[X|Z_i]$ as

$$h_P(Z_i) \left(\sum_{l \in \mathcal{M}} h_P(z_l) h_P(z_l)' \right)^{-1} \sum_{l \in \mathcal{M}} h_P(z_l) X_l,$$

and calculate the deviation from the expectation as

$$\tilde{X}_i = X_i - \hat{\mathbb{E}}[X|Z_i] = X_i - h_P(Z_i) \left(\sum_{l \in \mathcal{M}} h_P(z_l) h_P(z_l)' \right)^{-1} \sum_{l \in \mathcal{M}} h_P(z_l) X_l.$$

Similarly we estimate the deviation from its mean for the outcome as

$$\tilde{Y}_i = Y_i - \hat{\mathbb{E}}[Y|Z_i] = Y_i - h_P(Z_i)\left(\sum_{l\in\mathcal{M}} h_P(z_l)h_P(z_l)'\right)^{-1}\sum_{l\in\mathcal{M}} h_P(z_l)Y_l.$$

Finally, we estimate β as

$$\hat{\beta} = \left(\sum_{i=1}^{N}\tilde{X}_i\tilde{X}_i'\right)^{-1}\sum_{i=1}^{N}\tilde{X}_i\tilde{Y}_i.$$

Theorem 4.1 ()
Suppose Then (i):

$$\hat{\beta} - \beta = O_p(N^{-1/2}),$$

and (ii),

$$\sqrt{N}\hat{V}^{-1/2}(\hat{\beta}-\beta) \xrightarrow{d} \mathcal{N}(0, I_{\ L}).$$

5 Application: Robinson's Partial Linear Model

We use a subsample of the 1979 National Longitudinal Survey of Youth (NLSY), with observations on earnings, education, age, and two measures of ability, an iq test and a testscore referred to as kww (knowledge of the world of work). Means and standard devations for the sample are given in Table 1. In this table the variable experience is calculated as age minus six minus years of education.

We consider the following model:

$$\log(\text{earn})_i = \beta_1\text{educ}_i + \beta_2\text{exper}_i + \beta_3\text{exper}_i^2 + g(\text{iq}_i, \text{kww}_i) + \varepsilon_i, \tag{5.6}$$

where $\mathbb{E}[\varepsilon_i|\text{educ}, \text{exper}, \text{iq}, \text{kww}] = 0$.

The parameter of interest is the returns to education, that is, the coefficient on years of education, β_2. We consider seven estimators for this parameter. The first is based on a linear regression model with no controls for iq and kww:

$$\log(\text{earn})_i = \beta_1\text{educ}_i + \beta_2\text{exper}_i + \beta_3\text{exper}_i^2 + \varepsilon_i.$$

Next, we add linear controls for iq and kww:

$$\log(\text{earn})_i = \beta_1\text{educ}_i + \beta_2\text{exper}_i + \beta_3\text{exper}_i^2 + \beta_4\text{iq}_i + \beta_5\text{kww}_i + \varepsilon_i.$$

The third estimator adds second order terms for iq and kww:

$$\log(\text{earn})_i = \beta_1\text{educ}_i + \beta_2\text{exper}_i + \beta_3\text{exper}_i^2 + \beta_4\text{iq}_i + \beta_5\text{kww}_i + \beta_6\text{iq}_i^2 + \beta_7\text{kww}_i^2 + \beta_8\text{iq}_i\text{kww}_i + \varepsilon_i.$$

The next five estimators are based on first estimating residuals from regression the outcome variable $Y = \log(\text{earn})$ and covariates $X = (\text{educ}, \text{exper}, \text{exper}^2)$ on $Z = (\text{iq}, \text{kww})$. The remaining five estimators differ in the way they estimate the conditional expectation given Z. Let us focus on estimation of the conditional expectation of Y given Z, since estimation of the conditional expectation of each component of X given Z works in the same way. The fourth estimator uses the value of Y for the single nearest match given Z. The fifth estimator uses the three nearest matches and averages the Y for those three matches. The sixth estimator uses the three nearest matches, and then regression adjusts for the differences in Z between the matches and the matched observations using linear regression. The seventh and eight are similar to the fifth and sixth in that they use three matches, but they differ in the way the three matches are choosen. Instead of picking the three closest matches, the three matches are choosen to be the closest subject subject to the constraint that the maximum weight for the three matches in predicting the outcome for the unit that is being matches is restricted to be less than or equal to 3.

Results for estimating this on the NLS data set are given in Table 2.

6 Simulations

For the simulations we first estimate the linear regression model with second order terms on the NLS data with 935 observations:

$$\ln \text{earn}_i = \beta_1 \text{educ}_i + \beta_2 \text{exper}_i + \beta_3 \text{exper}_i^2 + \beta_4 \text{iq}_i + \beta_5 \text{kww}_i + \beta_6 \text{iq}_i^2 + \beta_7 \text{kww}_i^2 + \beta_8 \text{iq}_i \text{kww}_i + \varepsilon_i, \quad (6.7)$$

and $\mathbb{E}[\varepsilon_i | \text{educ}_i, \text{exper}_i, \text{iq}_i, \text{kww}_i] = \sigma^2$. The estimates for the parameters are presented in Tabel 3. The variables iq and kww are normalized by subtracting 100 and 36 respectively, and dividing by 10 and 7.

For the simulations we use two data generating processes. In both cases the joint distribution of educ, exper, iq and kww is normal with the vector of means and the covariance matrix estimated from the NLS data (after iq is transformed by subtracting 10 and dividing by 10 and kww is transformed by subtracting 36 and dividing by 7. The means, standard deviation and correlation matrix are given in Tabel 1. After generating these variables exper^2 is generated as the square of exper. Finally, the outcome log(earn) is generated using the model with second order terms in iq and kww given in (6.7) and normal residuals. The first set of parameter values is very similar to those estimated on the NLS data. The one modification is that the returns to education are set equal to 0.06. The second set of parameter values differ from the first only in the coefficients on the five iq and kww terms which are all multiplied by 20 to increase the biases from failing to adjust for the iq and kww covariates. These two sets of parameter values are presented in Tabel 3. In the simulations we use two different sample sizes, $N = 100$ and $N = 800$. Results (mean and median bias, root mean squared error, and median absolute error) for the simulations are presented in Table 4

7 Conclusion

Appendix

Lemma A.1 (Existence)
For any P and K, there is a set $\{x_1, \ldots, x_{M(P,K)}\}$ of elements of $\mathbb{R}^K$ such that the $M(P,K) \times M(P,K)$ matrix with m^{th} column equal to $h_P(x_m)$ is nonsingular.

Proof of Lemma A.1:
Fix K. For any positive integer j, let $p = \min\{p' | M(p', K) \geq j\}$. Then define $h_p^{1:j}(x)$ as the first j elements of $h_p(x)$ where the elements of $h_p(x)$ are ordered such that for K-dimensional vectors of nonnegative integers λ, λ' if $|\lambda| < |\lambda'|$ then x^λ is listed before $x^{\lambda'}$ in the vector $h_P(x)$. Further $h_p^j(x)$ will denote the i^{th} element of the vector $h_p(x)$.

We will show the following: For any j and K, there is a set $\{x_1, \ldots, x_j\}$ of elements of $\mathbb{R}^K$ such that the $j \times j$ matrix $[h_p^{1:j}(x_1) \cdots h_p^{1:j}(x_j)]$ is nonsingular. This assertion is stronger than the statement of the lemma. We prove it by induction.

First, consider the case $j = 1$, any K. Let x be a nonzero element of $\mathbb{R}^K$. $h_p^1(x) = 1$ is nonsingular. Hence the result holds for $j = 1$, and any K.

Second, consider the case $j = 2$, any K. Choose $x_1 \in \mathbb{R}^K$ to be any vector with first element 2 and $x_2 \in \mathbb{R}^K$ any vector with first element 3, ie $x_1 = \begin{pmatrix} 2 \\ \vdots \end{pmatrix}$, and $x_2 = \begin{pmatrix} 3 \\ \vdots \end{pmatrix}$. Then $[h_p^{1:2}(x_1) h_p^{1:2}(x_2)] = \begin{pmatrix} 1 & 1 \\ 2 & 3 \end{pmatrix}$ is nonsingular. Hence the result holds for $j = 2$ and any K.

Now assume that the result holds for up to j and any K. Then there exists a set $\{x_1, \ldots, x_j\}$ of elements of $\mathbb{R}^K$ such that $[h^{K,j}(x_1) \cdots h^{K,j}(x_j)]$ is nonsingular. We will show that there exists a $z \in \mathbb{R}^K$ such that $[h^{K,j+1}(x_1) \cdots h^{K,j+1}(x_j) h^{K,j+1}(z)]$ is nonsingular.

Note that $\text{rank}([h^{K,j}(x_1) \cdots h_p^{1:j}(x_j)]) = j$, so $\text{rank}([h_p^{1:j+1}(x_1) \cdots h_p^{1:j+1}(x_j)]) = j$. Given $z \in \mathbb{R}^K$, $[h_p^{1:j+1}(x_1) \cdots h_p^{1:j+1}(x_j) h_p^{1:j+1}(z)]$ is nonsingular if and only if the only solution for $a \in \mathbb{R}^j$ to

$$h_p^{1:,j+1}(z) - [h_p^{1:j+1}(x_1) \cdots h_p^{1:j+1}(x_j)]a = 0$$

is $a = 0$. There exists a unique nonzero $a \in \mathbb{R}^j$ such that

$$h_p^{1:j}(z) - [h_p^{1:j}(x_1) \cdots h_p^{1:j}(x_j)]a = 0.$$

This unique solution is

$$a = [h_p^{1:j}(x_1) \cdots h_p^{1:j}(x_j)]^{-1} h_p^{1:j}(z).$$

Now let

$$l(z) = h_p^{j+1}(z) - [h_p^{j+1}(x_1) \cdots h_p^{j+1}(x_j)][h_p^{1:j}(x_1) \cdots h_p^{1:j}(x_j)]^{-1} h_p^{1:j}(z)$$

Then $[h_p^{1:j+1}(x_1) \cdots h_p^{1:j+1}(x_j) h^{1:j+1}(z)]$ is nonsingular if and only if $l(z) \neq 0$. That is, the lemma will follow if we can show $l(z)$ is not identically zero. For some λ, $h_p^{j+1}(z) = z^\lambda$. For this λ, $\frac{d^\lambda}{dz^\lambda} h_p^i(z) = 0$ for $i = 1, \ldots j$ and $\frac{d^\lambda}{dz^\lambda} h_p^{j+1}(z) = 1$. Hence, $\frac{d^\lambda}{dz^\lambda} l(z) = 1$, so continuity of $l(z)$ implies that $l(z)$ cannot be identically zero. □

Proof of Lemma 3.1:

Let $x_1, \ldots, x_{M(P,K)}$ be a set of points satisfying Lemma A.1. Then we can also find a set of points satisfying this condition inside the unit ball $\mathbb{B}_K = \{x \in \mathbb{R}^K | \|x\| \leq 1\}$ by multiplying each point x_i by $c = \left(\max_{i=1,\ldots,M(P,K)} \|x_i\|\right)^{-1}$. (This changes the value of the function x^λ by $c^{|\lambda|}$, and so the determinant by $c^{M(P,K)\cdot|\lambda|}$, and therefore does not change the nonsingularity of the matrix.) Hence there is a set of points $x_1, \ldots, x_{M(P,K)}$ in the ball $\mathbb{B}_K$ such that the $M(P,K) \times M(P,K)$ matrix with mth column equal to $h_P(x_m)$ is nonsingular and positive definite. For this set of points the corresponding $H = \sum_{m=1}^{M(P,K)} h_P(x_m) h_P(x_m)'/M(P,K)$ is nonsingular, so the eigenvalues of H are bounded away from zero, hence the eigenvalues of H^{-1} are bounded. Let ζ_{max} denote the maximum eigenvalue of H^{-1}. Using a matrix version of the Cauchy-Schwarz inequality for nonnegative definite matrices, for any $x, z \in \mathbb{B}_K$,

$$\begin{aligned}
h_P(x)'H^{-1}h_P(z) &\leq \sqrt{h_P(x)'H^{-1}h_P(x)}\sqrt{h_P(z)'H^{-1}h_P(z)} \\
&\leq \sqrt{\frac{h_P(x)'H^{-1}h_P(x)}{h_P(x)'h_P(x)}}\sqrt{\frac{h_P(z)'H^{-1}h_P(z)}{h_P(z)'h_P(z)}} \sup_{w \in \mathbb{B}_K} h_p(w)'h_P(w) \\
&\leq \zeta_{max} \cdot \sup_{w \in \mathbb{B}_K} h_p(w)'h_P(w) \\
&\leq \zeta_{max} \cdot M(P,K) \\
&< \infty
\end{aligned}$$

where the fourth inequality follows by noting that

$$\sup_{w \in \mathbb{B}_K} h_p(w)'h_P(w) = \sup_{w \in B_K} \sum_{\lambda \in \Lambda_P} (w^\lambda)^2 \leq \sum_{\lambda \in \Lambda_P} \sup_{w \in B_K} (w^\lambda)^2 \leq \sum_{\lambda \in \Lambda_P} 1 = M(P,K).$$

Therefore the supremum over $z \in \mathbb{B}_K$ and the maximum over $m = 1, \ldots, M(P,K)$ of $h_P(z)'H^{-1}h_P(x_m)$ for such a set of points is finite. □

Before proving Theorem 3.1 we state and prove a couple of preliminary results.

Lemma A.2 (POSITIVE PROBABILITY)

Let $f_X(x)$ be a density on $\mathbb{B}_K$, bounded from below by $\underline{f} > 0$. For any $\varepsilon > 0$, there is a $\underline{p}(P,K) > 0$ such that for a random sample $X_1, \ldots, X_{M(P,K)}$ of size $M(P,K)$ from this distribution,

$$\Pr\left(\sup_{z \in \mathbb{B}_K} \max_{m \in \{1,2,\ldots,M(P,K)\}} h_P(z)'H(X_1, \ldots, X_{M(P,K)})^{-1}h_P(X_m) < \delta(P,K) + \varepsilon\right) \geq \underline{p}(P,K). \quad \text{(A.1)}$$

(if the matrix H is singular, we define $\sup_{z \in B_K} h_P(z)'H^{-1}h_P(x)$ to be infinite.)

We index the bound by P and K to stress that it depends on the values of these parameters. The bound also depends on the lower bound of the density $\underline{f}$.

Proof of Lemma A.2:

By Lemma 3.1 there is a set of $\mathbb{M} = \{x_1, \ldots, x_{M(P,K)}\} \in \mathbb{B}_K$ such that

$$\sup_{z \in \mathbb{B}_K} \max_{x \in \mathbb{M}} h_P(z)'H(\mathbb{M})^{-1}h_P(x) = \delta(P,K).$$

Hence, by continuity of the inverse of a nonsingular matrix in its elements, it follows that there exists an $\eta > 0$ such that for all $x'_1, \ldots, x'_{M(P,K)}$ with $\|x'_m - x_m\| \leq \eta$, for $m = 1, \ldots, M(P,K)$, we have

$$\sup_{z \in \mathbb{B}_K} \max_{m=1,\ldots,M(P,K)} h_P(z)H(x'_1, \ldots, x'_{M(P,K)})^{-1}h_P(x'_m) < \delta(P,K) + \varepsilon.$$

Then a lower bound on the lefthand side of (A.1) is

$$\underline{p} = \left(\frac{\underline{f} \eta^K \pi^{K/2}}{\Gamma(1+K/2)} \right)^{M(P,K)},$$

since

$$\begin{aligned} &\Pr\left(\sup_{z \in B_K} \max_m h_P(z)' H(X_1, \ldots, X_{M(P,K)})^{-1} h_P(X_m) < \delta(P,K) + \varepsilon \right) \\ &\geq \ Pr\left(\cap_{m=1}^{M(P,K)} \|X_m - x_m\| \leq \eta \right) \\ &= \ \prod_{m=1}^{M(P,K)} Pr(\|X - x_m\| \leq \eta) \quad \text{(by independence)} \\ &\geq \ \left(\frac{\underline{f} \eta^K \pi^{K/2}}{\Gamma(1+K/2)} \right)^{M(P,K)}, \end{aligned}$$

where $\underline{f}$ is the lower bound on the density $f_X(x)$ and we use the fact that the volume of the K-dimensional unit ball is $\pi^{K/2}/\Gamma(1+K/2)$. □

Now consider a convex, compact subset $\mathbb{X}$ of $\mathbb{R}^K$ with non-empty interior, and density $f_X(x)$ bounded away from zero. Fix $0 < \alpha < 1/K$. Consider for each $z \in \mathbb{X}$ the ball $\mathbb{B}_{N,\alpha}(z) = \{x \in \mathbb{X} | \|x - z\| \leq N^{-\alpha}\}$. Let $X_1, \ldots, X_N$ be a random sample from density $f_X(x)$ (with $f_X(x) \geq \underline{f}$ for $x \in \mathbb{X}$ and $f_X(x) = 0$ for $x \notin \mathbb{X}$), and let $\mathbb{M} = \{x_1, \ldots, x_{M(P,K)}\}$ be a subset of $\mathbb{R}^K$ with $M(P,K)$ elements. Define the probability

$$p_N(z) = \Pr\left(\exists \mathbb{M} \middle| \mathbb{M} \subset \{X_1, \ldots, XN\}, \mathbb{M} \subset \mathbb{B}_{N,\alpha}(z), \sup_{y \in \mathbb{B}_{N,\alpha}(z)} \max_{x \in \mathbb{M}} h_P(y)' H(\mathbb{M})^{-1} h_P(x) < \delta(P,K) + \varepsilon \right).$$

Also define $N_{\mathbb{B}}(z)$ to be the cardinality of the intersection of $\mathbb{B}_{N,\alpha}(z)$ and $\{X_1, \ldots, X_N\}$:

$$N_{\mathbb{B}} = \sum_{i=1}^{N} 1\left\{ X_i \in \mathbb{B}_{K,N^{-\alpha}}(z) \right\}.$$

Lemma A.3 (The Number of Close Matches)
For all $z \in$ int $\mathbb{X}$, (i):

$$\Pr(N_B(z) \leq M) \leq \exp(-CN^{1-\alpha K}) \cdot M \cdot C^M \cdot N^{(1-\alpha K)M},$$

for $C = \underline{f}\pi^{(K/2)}/\Gamma(1+K/2)$, and for all $\epsilon > 0$, (ii):

$$\lim_{N \to \infty} \Pr(N_B(z) \geq N^{1-\alpha K - \varepsilon}) \to 1.$$

Proof of Lemma A.3: The probability that X_i is in the set $B_{N,\alpha}(z)$ is bounded from below by $\underline{f}$ times the volume of the ball. The volume of the ball $B_{N,\alpha}(z)$ with radius $N^{-\alpha}$ in $\mathbb{R}^K$ is $N^{-\alpha K}$ times the volume of a unit ball in $\mathbb{R}^K$, which is $\pi^{K/2}/\Gamma(1+K/2)$. Hence the volume of the ball $B_{N,\alpha}(z)$ is $N^{-\alpha K}\pi^{(K/2)}/\Gamma(1+K/2)$, and the probability that X_i is in the set $\mathbb{B}_{N,\alpha}(z)$ is bounded from below by $p = \underline{f}N^{-\alpha K}\pi^{(K/2)}/\Gamma(1+K/2) = CN^{-\alpha K}$ for $C = \underline{f}\pi^{(K/2)}/\Gamma(1+K/2) > 0$. Since the number of observations $N_B(z)$ that fall inside the ball has a binomial distribution with parameters N and $p = \Pr(X_i \in \mathbb{B}_{N,\alpha}(z))$, it follows that for all M,

$$\Pr(N_b(z) \leq M) \leq \sum_{m=0}^{M} \begin{pmatrix} N \\ m \end{pmatrix} p^m \, (1-p)^{N-m} = \sum_{m=0}^{M} \frac{N!}{m!(N-m)!} p^m (1-p)^{N-m}$$

$$\leq \sum_{m=0}^{M} \frac{N^m}{m!} p^m (1-p)^{N-m} = \sum_{m=0}^{M} \frac{1}{m!} \tilde{p}^m (1-\tilde{p}/N)^{N-m},$$

where $\tilde{p} = p \cdot N = CN^{1-\alpha K}$. Using the fact that $(1-a/N)^N \leq \exp(-a)$ we can further bound this from above by

$$\sum_{m=0}^{M} \frac{1}{m!} \tilde{p}^m (1-\tilde{p}/N)^{-m} \cdot \exp(-\tilde{p}) \leq \sum_{m=0}^{M} \frac{1}{m!} \cdot \tilde{p}^m \cdot \exp(-\tilde{p})$$

$$\leq M \cdot \tilde{p}^M \cdot \exp(-\tilde{p})$$

$$= \exp(-CN^{1-\alpha K}) \cdot M \cdot C^M \cdot N^{(1-\alpha K)M}. \tag{A.2}$$

which gives the first result.

For the second result, take the log of the first term on the right hand side of (A.2) to get

$$-CN^{1-\alpha K} + \ln M + M \ln C + (1-\alpha K)M \cdot \ln N.$$

Substitute $M = N^{1-\alpha K-\epsilon}$ to get

$$-CN^{1-\alpha K} + (1-\alpha K-\epsilon) \cdot \ln N + \ln C \cdot N^{1-\alpha K-\epsilon} + \ln N \cdot (1-\alpha K) \cdot N^{1-\alpha K-\epsilon}$$

$$= N^{1-\alpha K-\epsilon} \cdot \Big(-C \cdot N^{\epsilon} + N^{-(1-\alpha K-\epsilon)}(1-\alpha K-\epsilon) \cdot \ln N + \ln C + \ln N \cdot (1-\alpha K) \Big).$$

The second factor is dominated by the term $-CN^{\epsilon}$. Since $N^{1-\alpha K}$ goes to infinity, the product goes to minus infinity. Since this was an upper bound on the log of the probability $\Pr(N_B(z) \leq N^{1-\alpha K-\varepsilon})$, this probability converges to zero, implying the second result in the Lemma. □

Lemma A.4 (PROBABILITY OF EXISTENCE OF SUITABLE SET OF MATCHES GOES TO ONE)

$$\lim_{N\to\infty} p_N(z) = 1.$$

Proof of Lemma A.4: First let us consider the probability that a random sample of size $M(P,K)$ drawn from the set $\mathbb{B}_{N,\alpha}(z)$, satisfies the condition

$$\sup_{y \in \mathbb{B}_{N,\alpha}(z)} \max_m h_P(y)' H(X_1, \ldots, X_{M(P,K)})^{-1} h_P(X_m) < \delta(P,K) + \varepsilon. \tag{A.3}$$

Note that if the original density $f_X(x)$ is bounded and bounded away from zero, then the conditional density within the set $\mathbb{B}_{N,\alpha}(z)$ is also bounded and bounded away from zero. Also note that $h_P(cx) = C \cdot h_p(X)$ for a diagonal matrix C, with ith diagonal element equal to $c^{|\lambda_i|}$ if the ith function $h_{P,i}(x) = x^{\lambda_i}$. Hence

$$h_P(cy)' H(cX_1, \ldots, cX_{M(P,K)})^{-1} h_P(cX_m) = h_P(y) C \left(CH(X_1, \ldots, X_{M(P,K)}) C \right)^{-1} C h_P(X_m)$$

$$= h_P(y)'H(X_1,\ldots,X_{M(P,K)})^{-1}h_P(X_m).$$

Hence condition (A.3) is the same as

$$\sup_{y\in\mathbb{B}_{N,\alpha}(z)}\max_m h_P(N^\alpha y)'H(N^\alpha X_1,\ldots,N^\alpha X_{M(P,K)})^{-1}h_P(N^\alpha X_m)<\delta(P,K)+\varepsilon.$$

Because $y\in\mathbb{B}_{N,\alpha}(z)$ is equivalent to $N^\alpha(y-z)\in\mathbb{B}_K$, this is the same as the condition

$$\sup_{y\in\mathbb{B}_K}\max_m h_P(y)'H(N^\alpha X_1,\ldots,N^\alpha X_{M(P,K)})^{-1}h_P(N^\alpha X_m)<\delta(P,K)+\varepsilon.$$

Because the support of X is $\mathbb{B}_{N,\alpha}(z)$, the support of $N^\alpha(X-z)$ is $\mathbb{B}_K$, so the probability of this condition being satisfied is greater than $\underline{p}$ by Lemma A.2, where $\underline{p}$ is computed with respect to the density of $N^\alpha X$ on $\mathbb{B}_K$, which is bounded and bounded away from zero.

With $N_{\mathbb{B}}(z)$ is the number of points inside the ball $\mathbb{B}_{N,\alpha}(z)$ there are $N_{\mathbb{B}}(z)/M(P,K)$ non-overlapping sets $\mathbb{M}$. Therefore, with the probability of any one set satisfying the condition bounded from below by $\underline{p}$, the probability of the existence of at least one such set (in a set of $N_{\mathbb{B}}(z)/M(P,K)$ sets) that satisfies the condition is bounded from below by $1-(1-\underline{p})^{N_B/M(P,K)}$. Given $\eta>0$, choose $M>M(P,K)\cdot\ln(\eta/2)/\ln(1-\underline{p})$ so that $(1-\underline{p})^{M/M(P,K)}<\frac{\eta}{2}$ and choose N large enough that $Pr(N_{\mathbb{B}}(z)<M)<\eta/2$ (this is possible because Lemma A.3). Then,

$$\begin{aligned}
p_N(z) &= \Pr\left(\exists\mathbb{M}\subset\mathbb{B}_{N,\alpha}(z),\ \sup_{y\in\mathbb{B}_{N,\alpha}(z)}\max_{x\in\mathbb{M}}h_P(y)'H(\mathbb{M})^{-1}h_P(x)<\delta(P,K)+\varepsilon\right)\\
&= \Pr\left(\exists\mathbb{M}\subset\mathbb{B}_{N,\alpha}(z),\ \sup_{y\in\mathbb{B}_{N,\alpha}(z)}\max_{x\in\mathbb{M}}h_P(y)'H(\mathbb{M})^{-1}h_P(x)<\delta(P,K)+\varepsilon\,\middle|\,N_B\ge M\right)Pr(N_B\ge M)\\
&\quad+\Pr\left(\left(\exists\mathbb{M}\subset\mathbb{B}_{N,\alpha}(z),\ \sup_{y\in\mathbb{B}_{N,\alpha}(z)}\max_{x\in\mathbb{M}}h_P(y)'H(\mathbb{M})^{-1}h_P(x)<\delta(P,K)+\varepsilon\right)\cap(\{N_B<M\})\right)\\
&\ge Pr\left(\exists\mathbb{M}\subset\mathbb{B}_{N,\alpha}(z),\ \sup_{y\in\mathbb{B}_{N,\alpha}(z)}\max_{x\in\mathbb{M}}h_P(y)'H(\mathbb{M})^{-1}h_P(x)<\delta(P,K)+\varepsilon\,\middle|\,N_B\ge M\right)\left(1-\frac{\eta}{2}\right)\\
&\ge \left(1-(1-\underline{p})^{M/M(P,K)}\right)\left(1-\frac{\eta}{2}\right)\\
&\ge \left(1-\frac{\eta}{2}\right)\left(1-\frac{\eta}{2}\right)\\
&\ge 1-\eta.
\end{aligned}$$

The conclusion of the lemma follows. □

Proof of Theorem 3.1:

Define $\mathbb{B}^\nu_K(z)=\{x\in\mathbb{R}^K|\|x-z\|\le\nu\}$. For some finite L_1 we can cover the set $\mathbb{X}$ with a set of L_1 unit balls of the form $\mathbb{B}^1_K(z_l)$, for $l=1,\ldots,L$. For any $0<\nu<1$ there is a finite L_ν such that we can cover the ball $\mathbb{B}^1_K(z)$ by L_ν balls of the form $\mathbb{B}^\nu_K(z_l)$. By the same argument, each of those can in turn be covered by L_ν balls of the form $\mathbb{B}^{\nu^2}_K(z_l)$. Hence we can cover the unit ball $B^1_K(z)$ by L^R_ν balls of the form $B^{\nu^R}_K(z_l)$. Therefore, we can cover the unit ball $B^1_K(z)$ by $(L_\nu)^R$ balls of the form $B^{N^{-\alpha}}_K(z_l)$ if $\nu^R\le N^{-\alpha}$. This implies that in order to can cover the unit ball $B^1_K(z)$ it is sufficient that $R\ge 1-\alpha\cdot\ln N/\ln\nu$, and thus that we need at most $(L_\nu)^{1-\alpha\ln N/\ln\nu}=N^{-\alpha\ln L_\nu/\ln\nu}\cdot L_\nu$ balls of the form $\mathbb{B}^{N^{-\alpha}}_K(z_l)$. To cover $\mathbb{X}$ we therefore need at most $N^{-\alpha\ln L_\nu/\ln\nu}\cdot L_\nu\cdot L_1=C_1\cdot N^{\alpha\cdot C_2}$ balls $\mathbb{B}^{N^{-\alpha}}_K(z_l)$.

Let N_l be the number of observations in ball $\mathbb{B}^{N^{-\alpha}}_K(z_l)$ for $l=1,\ldots,C_1\cdot N^{\alpha\cdot C_2}$. For fixed $\epsilon>0$ we want to bound the probability that for any l, $N_l\le N^{1-\alpha K-\epsilon}$. For fixed l we have (using the first part of Lemma A.3 and substituting $M = N^{1-\alpha K-\epsilon}$)

$$\Pr(N_l \leq N^{1-\alpha K-\epsilon})$$

$$\leq \exp(-CN^{1-\alpha K})N^{1-\alpha K-\epsilon} \cdot \exp(\ln C \cdot N^{1-\alpha K-\epsilon}) \cdot \exp(\ln N \cdot (1-\alpha K)N^{1-\alpha K-\epsilon}),$$

for some positive C. To bound the probability that for any of the sets the number of observations is less than $N^{1-\alpha K-\epsilon}$ we multiply the probability for any one set by the number of sets, $C_1 \cdot N^{\alpha \cdot C_2}$ to get;

$$\Pr(\exists l \text{ s.t. } N_l \leq N^{1-\alpha K-\epsilon})$$

$$\leq C_1 N^{C_2} \cdot \exp(-CN^{1-\alpha K})N^{1-\alpha K-\epsilon} \cdot \exp(\ln C \cdot N^{1-\alpha K-\epsilon}) \cdot \exp(\ln N \cdot (1-\alpha K)N^{1-\alpha K-\epsilon}).$$

Taking logs we get

$$\ln C_1 + (1-\alpha K) \cdot \ln N - CN^{1-\alpha K} + (1-\alpha K-\epsilon) \cdot \ln N + \ln C \cdot N^{1-\alpha K-\epsilon} + \ln N \cdot (1-\alpha K) \cdot N^{1-\alpha K-\epsilon}$$

$$= N^{1-\alpha K-\epsilon} \cdot \Big(\ln C_1 \cdot N^{-(1-\alpha K-\epsilon)} + (1-\alpha K) \cdot \ln N \cdot N^{-(1-\alpha K-\epsilon)} - C \cdot N^{\epsilon} +$$

$$N^{-(1-\alpha K-\epsilon)}(1-\alpha K-\epsilon) \cdot \ln N + \ln C + \ln N \cdot (1-\alpha K)\Big).$$

The second factor is dominated by the term $-C \cdot N^{\epsilon}$, so that the entire expression converges to zero. Hence we can find for any $\eta > 0$ an $\underline{N}$ such that the probability $\Pr(\exists l \text{ s.t. } N_l \leq N^{1-\alpha K-\epsilon})$ is less than $\eta/2$.

Conditional on $N_l > N^{1-\alpha K-\epsilon}$ for all l, the probability of at least one ball with no suitable set of matches (that is no set with the weights limited as in Lemma A.2) is bounded from above by

$$C_1 N^{C_2}(1-\underline{p}(P,K))^{N^{1-\alpha K-\epsilon}/M(P,K)},$$

by Lemma A.2. For N large enough we can make this probability less than $\eta/2$. For such N the probability of a suitable set of matches in each ball is therefore at least $1-\eta$. □

Proof of Lemma 3.2:

Let h_P be ordered as described in the proof of Lemma A.1. If $h_{P,j}(x)$ is the j^{th} element of the vector, then let $\lambda_{P,j}$ be the K-dimensional vector of nonnegative integers such that $h_{P,j}(x) = x^{\lambda_{P,j}}$. Also, let $h^P(x)$ be a vector consisting of the last $M(P,K) - M(P-1,K)$ elements of $h_P(x)$, ie the elements x^λ with $|\lambda| = P$.

Now a Taylor expansion (with remainder) of μ to order $P+1$ around x can be expressed as follows:

$$\mu(z) = h_P(z-x)' \begin{pmatrix} \mu(x) \\ \nabla^{\lambda_{P,2}}\mu(x) \\ \vdots \\ \nabla^{\lambda_{P,M(P,K)}}\mu(x) \end{pmatrix} + h^{P+1}(z-x)' \begin{pmatrix} \nabla^{\lambda_{P+1,M(P,K)+1}}\mu(\bar{z}_x) \\ \vdots \\ \nabla^{\lambda_{P,M(P+1,K)}}\mu(\bar{z}_x) \end{pmatrix}$$

where $\nabla^\lambda \mu(x) = \frac{\partial^{|\lambda|}\mu}{\partial x^\lambda}(x)$.

Now note that for all z, there exists a lower triangular matrix $A(x)$ such that $h_P(z-x) = A(x)h_P(z)$. Further all of $A(x)$'s diagonal elements are equal to one, so $A(x)$ is nonsingular. For example, for

$K = P = 2$,

$$\begin{pmatrix} 1 \\ z_1 - x_1 \\ z_2 - x_2 \\ (z_1 - x_1)^2 \\ (z_2 - x_2)^2 \\ (z_1 - x_1)(z_2 - x_2) \end{pmatrix} = \begin{pmatrix} 1 \\ z_1 - x_1 \\ z_2 - x_2 \\ z_1^2 - 2z_1x_1 + x_1r \\ z_2^2 - 2z_2x_2 + x_2^2 \\ z_1z_2 - x_1z_2 - x_2z_1 + x_1x_2 \end{pmatrix} = \begin{pmatrix} 1 & 0 & 0 & 0 & 0 & 0 \\ -x_1 & 1 & 0 & 0 & 0 & 0 \\ -x_2 & 0 & 1 & 0 & 0 & 0 \\ x_1^2 & -2x_1 & 0 & 1 & 0 & 0 \\ x_2^2 & 0 & -2x_2 & 0 & 1 & 0 \\ x_1x_2 & -x_2 & -x_1 & 0 & 0 & 1 \end{pmatrix} \begin{pmatrix} 1 \\ z_1 \\ z_2 \\ z_1^2 \\ z_2^2 \\ z_1z_2 \end{pmatrix}.$$

Hence $h_P(z) = A(x)^{-1}h_P(z - x)$. Also if we set $z = x$ then $h_P(x) = A(x)^{-1}h_P(0) = A(x)^{-1}e_1$, where e_1 is a vector with first element equal to one and zeroes elsewhere.

$$\begin{aligned}
\hat{\mu}_{P,\mathbb{M}}(x) - \mu(x) &= \sum_{z\in\mathbb{M}_{M(P,K)}} \omega_z(x)\mu(z) - \mu(x) \\
&= \sum_{z\in\mathbb{M}_{M(P,K)}} h_P(x)' \left[\sum_{w\in\mathbb{M}_{M(P,K)}} h_P(w)h_P(w)'\right]^{-1} h_P(z)\mu(z) - \mu(x) \\
&= h_P(x)' \left[\sum_{z\in\mathbb{M}_{M(P,K)}} h_P(z)h_P(z)'\right]^{-1} \sum_{z\in\mathbb{M}_{M(P,K)}} h_P(z)\mu(z) - \mu(x) \\
&= {e_1}' A(x)^{-1\prime} \left[A(x)^{-1} \sum_{z\in\mathbb{M}_{M(P,K)}} h_P(z-x)h_P(z-x)' A(x)^{-1\prime}\right]^{-1} A(x)^{-1} \sum_{z\in\mathbb{M}_{M(P,K)}} h_P(z-x)\mu(z) - \mu(x) \\
&= {e_1}' \left[\sum_{z\in\mathbb{M}_{M(P,K)}} h_P(z-x)h_P(z-x)'\right]^{-1} \sum_{z\in\mathbb{M}_{M(P,K)}} h_P(z-x)\mu(z) - \mu(x) \\
&= {e_1}' \left[\sum_{z\in\mathbb{M}_{M(P,K)}} h_P(z-x)h_P(z-x)'\right]^{-1} \left\{ \sum_{z\in\mathbb{M}_{M(P,K)}} h_P(z-x)h_P(z-x)' \begin{pmatrix} \mu(x) \\ \nabla^{\lambda_{P,2}}\mu(x) \\ \vdots \\ \nabla^{\lambda_{P,M(P,K)}}\mu(x) \end{pmatrix} \right. \\
&\quad \left. + \sum_{z\in\mathbb{M}_{M(P,K)}} h_P(z-x)h^{P+1}(z-x)' \begin{pmatrix} \nabla^{\lambda_{P+1,M(P,K)+1}}\mu(\bar{z}_x) \\ \vdots \\ \nabla^{\lambda_{P,M(P+1,K)}}\mu(\bar{z}_x) \end{pmatrix} \right\} - \mu(x) \\
&= {e_1}' \left[\sum_{z\in\mathbb{M}_{M(P,K)}} h_P(z-x)h_P(z-x)'\right]^{-1} \sum_{z\in\mathbb{M}_{M(P,K)}} h_P(z-x)h^{P+1}(z-x)' \begin{pmatrix} \nabla^{\lambda_{P+1,M(P,K)+1}}\mu(\bar{z}_x) \\ \vdots \\ \nabla^{\lambda_{P,M(P+1,K)}}\mu(\bar{z}_x) \end{pmatrix} \\
&= h_P(x)'A(x)' \left[A(x) \sum_{z\in\mathbb{M}_{M(P,K)}} h_P(z)h_P(z)'A(x)'\right]^{-1} A(x) \\
&\quad \cdot \sum_{z\in\mathbb{M}_{M(P,K)}} h_P(z)h^{P+1}(z-x)' \begin{pmatrix} \nabla^{\lambda_{P+1,M(P,K)+1}}\mu(\bar{z}_x) \\ \vdots \\ \nabla^{\lambda_{P,M(P+1,K)}}\mu(\bar{z}_x) \end{pmatrix} \\
&= h_P(x)' \left[\sum_{z\in\mathbb{M}_{M(P,K)}} h_P(z)h_P(z)'\right]^{-1} \sum_{z\in\mathbb{M}_{M(P,K)}} h_P(z)h^{P+1}(z-x)' \begin{pmatrix} \nabla^{\lambda_{P+1,M(P,K)+1}}\mu(\bar{z}_x) \\ \vdots \\ \nabla^{\lambda_{P,M(P+1,K)}}\mu(\bar{z}_x) \end{pmatrix} \\
&= \sum_{z\in\mathbb{M}_{M(P,K)}} \omega_z(x)h^{P+1}(z-x)' \begin{pmatrix} \nabla^{\lambda_{P+1,M(P,K)+1}}\mu(\bar{z}_x) \\ \vdots \\ \nabla^{\lambda_{P,M(P+1,K)}}\mu(\bar{z}_x) \end{pmatrix}
\end{aligned}$$

For a vector of nonnegative integers λ, let

$$C_P = \max_{|lambda|=P} \sup_{x\in\mathbb{X}} \left|\nabla^{\lambda}\mu(x)\right|.$$

By Assumption 3.1 C_{P+1} is finite.

Hence, since $\mathbb{M} \subset\subset X$,

$$\sup_{x\in\mathbb{X}} |\hat{\mu}_{P,\mathbb{M}}(x) - \mu(x)| \leq \left[\sup_{z\in\mathbb{M}} \sup_{x\in\mathbb{X}} |\omega_z(x)|\right] M(P,K) C_{P+1} \sup_{x,z\in\mathbb{X}} \|x - z\|^{P+1}$$

which completest the proof. □

Proof of Theorem 3.2:
By Theorem 3.1,

$$Pr\left(\forall x \in \mathbb{X}, \mathbb{M}(x) \subset \mathbb{B}_{N,\alpha}(x)\right) \to 1.$$

For such x the bias is bounded by the supremum of the weights, which is $\delta(P,K)+\varepsilon$, times the supremum of the distances between points in $\mathbb{M}(x)$ and x, which is bounded by $N^{-\alpha}$, times the supremum of the $P+1$th derivative of $\mu(x)$, which is bounded by C_{P+1}. Hence the bias is bounded by $(\delta(P,K)+\varepsilon) \cdot C_{P+1} N^{-\alpha(P+1)}$.

REFERENCES

Abadie, A., and G. Imbens, (2002), "Simple and Bias-Corrected Matching Estimators for Average Treatment Effects," unpublished manuscript, Kennedy School of Government, Harvard University.

Estes, E.M., and B.E. Honoré, (2001), "Partially Linear Regression Using One Nearest Neighbor," unpublished manuscript, Princeton University.

Hahn, J., (1998), "On the Role of the Propensity Score in Efficient Semiparametric Estimation of Average Treatment Effects," *Econometrica* 66 (2), 315-331.

Heckman, J., H. Ichimura, and P. Todd, (1998), "Matching as an Econometric Evaluation Estimator," *Review of Economic Studies* 65, 261–294.

Hirano, K., G. Imbens, and G. Ridder, (2000), "Efficient Estimation of Average Treatment Effects Using the Estimated Propensity Score," NBER Working Paper.

Robinson, P., (1988), "Root-N-Consistent Semiparametric Regression," *Econometrica*, 67, 645-662.

Yatchew, A., (1999), "Differencing Methods in Nonparametric Regression: Simple Techniques for the Applied Econometrician", Working Paper, Department of Economics, University of Toronto.

Table 1: Summary Statistics NLS data

					Correlation Coefficients					
covariate	mean	s.d.	min	max	log(earn)	educ	age	exper	iq	kww
log(weekly earnings)	5.94	0.44	4.06	7.60	1.00	0.33	0.16	-0.06	0.33	0.33
years of education	13.47	2.20	9.00	18.00		1.00	-0.01	-0.58	0.52	0.39
age	33.08	3.11	28.00	38.00			1.00	0.82	-0.04	0.39
years of experience	13.61	3.83	5.00	23.00				1.00	-0.33	0.10
iq	101.28	15.05	50.00	145.00					1.00	0.41
kww	35.74	7.64	12.00	56.00						1.00

Table 2: Estimates for Returns to Education from NLS data

No controls	Linear	Quadratic	Matching Estimators				
			$M = 1$	$M = 3$ (Unrest)		$M = 3$ (Rest.)	
				unadj	adj	unadj	adj
0.0923	0.0560	0.0535	0.0490	0.0478	0.0267	0.0493	0.0445

Table 3: Estimates from NLS data and Parameter Values in Simulations

covariate	est.	s.e.	I	II
intercept	4.688	(0.249)	4.688	4.688
education	0.054	(0.01)	0.060	0.060
exper	0.060	(0.03)	0.060	0.060
exper^2	-0.002	(0.001)	-0.002	-0.002
iq	0.054	(0.011)	0.054	1.071
kww	0.066	(0.016)	0.066	1.312
iq^2	0.000	(0.005)	0.000	0.008
kww^2	0.028	(0.010)	0.028	0.561
iq×kww	-0.011	(0.011)	-0.011	-0.222
σ	0.398		0.398	0.398

Table 4: Mean Bias, Median Bias, RMSE and MAE (True Coefficient 0.06), Simulations with 10,000 Replications

design	sample size	No	Linear	Quadratic	$M = 1$	$M = 3$ (Unrest.)		$M = 3$ (Rest.)	
		Adj.	Adj.	Adj	Unadj.	Unadj.	Adj.	Unadj.	Adj.
I	100								
mean bias		0.0384	-0.0002	-0.0002	0.0007	0.0010	0.0026	0.0011	-0.0006
median bias		0.0386	-0.0001	-0.0000	0.0010	0.0010	0.0022	0.0008	-0.0007
rmse		0.0453	0.0289	0.0292	0.0394	0.0337	0.1831	0.0339	0.0509
mae		0.0386	0.0193	0.0196	0.0259	0.0220	0.0927	0.0223	0.0338
I	800								
mean bias		0.0390	0.0001	0.0001	0.0002	0.0003	-0.0000	0.0003	-0.0000
median bias		0.0390	0.0001	0.0001	0.0002	0.0003	0.0002	0.0003	0.0001
rmse		0.0398	0.0098	0.0098	0.0137	0.0118	0.0852	0.0119	0.0183
mae		0.0390	0.0067	0.0066	0.0092	0.0080	0.0516	0.0080	0.0122
II	100								
mean bias		0.7779	-0.0002	-0.0002	0.0178	0.0253	0.0026	0.0267	-0.0005
median bias		0.7749	-0.0000	-0.0000	0.0172	0.0233	0.0011	0.0247	-0.0007
rmse		0.7895	0.0632	0.0292	0.0532	0.0546	0.1840	0.0561	0.0514
mae		0.7749	0.0411	0.0196	0.0347	0.0348	0.0929	0.0359	0.0339
II	800								
mean bias		0.7779	0.0000	-0.0001	0.0029	0.0037	-0.0002	0.0037	0.0002
median bias		0.7772	-0.0001	-0.0001	0.0028	0.0036	0.0006	0.0037	0.0003
rmse		0.7793	0.0221	0.0097	0.0146	0.0133	0.0850	0.0134	0.0183
mae		0.7772	0.0150	0.0065	0.0098	0.0090	0.0521	0.0090	0.0123