Mean-squared-error Calculations for Average Treatment Effects *

Guido W. Imbens
UC Berkeley and NBER

Whitney Newey
MIT

Geert Ridder
USC

June 2007

Abstract

This paper develops a new nonparametric series estimator for the average treatment effect for the case with unconfounded treatment assignment, that is, where selection for treatment is on observables. The new estimator is efficient. In addition we develop an optimal procedure for choosing the smoothing parameter, the number of terms in the series by minimizing the mean squared error (MSE). The new estimator is linear in the first-stage nonparametric estimator. This simplifies the derivation of the MSE of the estimator as a function of the number of basis functions that is used in the first stage nonparametric regression. We propose an estimator for the MSE and show that in large samples minimization of this estimator is equivalent to minimization of the population MSE.

JEL Classification: C14, C20.

Keywords: *Nonparametric Estimation, Imputation, Mean Squared Error, Order Selection*

*We have benefitted from the comments of seminar participants at UCLA, Brown, Yale, UCL, and at the World Meetings of the Econometric Society in London. The research of Imbens and Ridder was supported by the National Science Foundation under grant SES-452572.

1 Introduction

Recently a number of estimators have been proposed for average treatment effects under the assumption of unconfoundedness or selection on observables. Many of these estimators require nonparametric estimation of an unknown function, either the regression function or the propensity score. Typically results are presented concerning the rates at which the smoothing parameters go to their limiting values, without specific recommendations regarding their values (Hahn, 1998; Hirano, Imbens and Ridder, 2003; Heckman, Ichimura and Todd, 1998; Rotnitzky and Robins, 1995; Robins, Rotnitzky and Zhao, 1995)).

In this paper we make two contributions. First, we propose a new efficient estimator. Our estimator is a modification of an estimator introduced in an influential paper by Hahn (1998). Like Hahn, our estimator relies on consistent estimation of the two regression functions followed by averaging their difference over the empirical distribution of the covariates. Our estimator differs from Hahn's in that it directly estimates these two regression functions whereas Hahn first estimates the propensity score and the two conditional expectations of the product of the outcome and the indicators for being in the control and treatment group, and then combines these to get estimates of the two regression functions. Thus our estimator completely avoids the need to estimate the propensity score.[1]

Our second and most important contribution is that we are explicit about the choice of smoothing parameters. In our series estimation setting the smoothing parameter is the number of terms in the series. We derive the population MSE of the imputation estimator which is simplified by the linearity of the imputation estimator in the nonparametrically estimated conditional expectations. We find that although the estimator does not require an estimate of the propensity score, the behavior of its MSE depends on the smoothness of the propensity score. In particular, the number of basis functions needed is determined by the smoothness of the propensity score, if the propensity score is smoother than the population regression of the outcome on the covariates. Hence, the regression estimator is undersmoothed. We show that our order selection criterion is different from the standard one for order selection in nonparametric regression as in Li (1987) and Andrews (1991), because our criterion focuses explicitly on optimal estimation of the average treatment effect rather than on optimal estimation of the entire unknown regression function. The population MSE has a squared bias term that must be estimated. We propose an estimator and we show that in large samples the minimal value of the population and estimated MSE are equal.

In the next section we discuss the basic set up and introduce the new estimator. In Section 3 we analyze the asymptotic properties of this estimator. In Section 4 we propose a method for choosing the number of terms in the series. Section 5 contains a simulation study and an empirical application.

[1] Independently Chen, Hong, and Tarozzi (2004) have established the efficiency of their CEP-GMM estimator which is similar to our new estimator.

2 The Basic Framework

The basic framework is standard in this literature (e.g Rosenbaum and Rubin, 1983; Hahn, 1998; Heckman, Ichimura and Todd, 1998; Hirano, Imbens and Ridder, 2003; Abadie and Imbens, 2005). We have a random sample of size N from a large population. For each unit i in the sample, let W_i indicate whether the treatment of interest was received, with $W_i = 1$ if unit i receives the treatment of interest, and $W_i = 0$ if unit i receives the control treatment. Using the potential outcome notation popularized by Rubin (1974), let $Y_i(0)$ denote the outcome for unit i under control and $Y_i(1)$ the outcome under treatment. We observe W_i and Y_i, where

$$Y_i \equiv Y_i(W_i) = W_i \cdot Y_i(1) + (1 - W_i) \cdot Y_i(0).$$

In addition, we observe a vector of pre-treatment variables, or covariates, denoted by X_i. We shall focus on the population average treatment effect:

$$\tau \equiv E[Y(1) - Y(0)].$$

Similar results can be obtained for the average effect for the treated:

$$\tau_t \equiv E[Y(1) - Y(0)|W = 1].$$

The central problem of evaluation research (e.g., Holland, 1986) is that for unit i we observe $Y_i(0)$ or $Y_i(1)$, but never both. Without further restrictions, the treatment effects are not consistently estimable. To solve the identification problem, we maintain throughout the paper the unconfoundedness assumption (Rubin, 1978; Rosenbaum and Rubin, 1983), which asserts that conditional on the pre-treatment variables, the treatment indicator is independent of the potential outcomes. This assumption is closely related to "selection on observables" assumptions (e.g., Barnow, Cain and Goldberger, 1980; Heckman and Robb, 1984). Formally:

Assumption 2.1 (Unconfoundedness)

$$W \perp (Y(0), Y(1)) \Big| X. \tag{2.1}$$

Let the propensity score be the probability of selection into the treatment group:

$$e(x) \equiv \Pr(W = 1|X = x) = \mathbb{E}[W|X = x], \tag{2.2}$$

The final assumption ensures that in the population for all values of X there are both treatment and control units.

Assumption 2.2 (Overlap)
The propensity score is bounded away from zero and one.

In practice there are often concerns about possible lack of overlap even if the assumption is formally satisfied. Crump, Hotz, Imbens and Mitnik (2005) propose a systematic method for constructing a subsample with an optimal amount of overlap.

Define the average effect conditional on pre-treatment variables:

$$\tau(x) \equiv \mathbb{E}[Y(1) - Y(0)|X = x]$$

Note that $\tau(x)$ is estimable under the unconfoundedness assumption, because

$$\begin{aligned}\mathbb{E}[Y(1) - Y(0)|X = x] &= \mathbb{E}[Y(1)|W = 1, X = x] - \mathbb{E}[Y(0)|W = 0, X = x] \\ &= \mathbb{E}[Y|W = 1, X = x] - \mathbb{E}[Y|W = 0, X = x].\end{aligned}$$

The population average treatment effect can be obtained by averaging the $\tau(x)$ over the distribution of X:

$$\tau = \mathbb{E}[\tau(X)],$$

and therefore the average treatment effect is identified.

3 Efficient Estimation

In this section we review two efficient estimators previously proposed in the literature. We then propose a new estimator. Different from the two existing efficient estimators, the new estimator is linear in the functions that are being estimated nonparametrically. This will facilitate the mean-squared error calculations.

3.1 The Hahn Estimator

Hahn (1998) studies the same model as in the current paper. He calculates the efficiency bound, and proposes an efficient estimator. His estimator imputes the potential outcomes given covariates, followed by averaging the difference of the imputed values. The difference with our estimator is that Hahn first estimates nonparametrically the three conditional expectations $\mathbb{E}[Y \cdot W|X]$, $\mathbb{E}[Y \cdot (1-W)|X]$, and $e(X) = \mathbb{E}[W|X]$. and then uses these conditional expectations to estimate the two regression function $\mu_0(x) = \mathbb{E}[Y(0)|X = x]$ and $\mu_1(x) = \mathbb{E}[Y(1)|X = x]$ as

$$\hat{\mu}_0(x) = \frac{\hat{\mathbb{E}}[Y(1 - W)|X = x]}{1 - \hat{e}(X)}, \qquad \text{and} \quad \hat{\mu}_1(x) = \frac{\hat{\mathbb{E}}[YW|X = x]}{\hat{e}(X)}.$$

The average treatment effect is then estimated as

$$\hat{\tau}_h = \frac{1}{N} \sum_{i=1}^{N} \left(\hat{\mu}_1(X_i) - \hat{\mu}_0(X_i) \right).$$

Hahn shows that under regularity conditions this estimator is consistent, asymptotically normally distributed, and that it reaches the semiparametric efficiency bound.

3.2 The Hirano-Imbens-Ridder Estimator

Hirano, Imbens and Ridder (2003) (HIR from here onwards) also study the same set up. They propose using a weighting estimator with the weights based on the estimated propensity score:

$$\hat{\tau}_{hir,1} = \frac{1}{N}\sum_{i=1}^{N} Y_i \cdot \left(\frac{W_i}{\hat{e}(X_i)} - \frac{1-W_i}{1-\hat{e}(X_i)}\right).$$

They show that under regularity conditions, and with $\hat{e}(x)$ a nonparametric estimator for the propensity score, this estimator is consistent, asymptotically normally distributed and efficient.

It will be useful to consider a slight modification of this estimator. Consider the weights for the treated observations, $1/\hat{e}(X_i)$. Summing up over all treated observations and dividing by N we get $\sum_i (W_i/\hat{e}(X_i))/N$. This is not necessarily equal to one. We may therefore wish to modify the weights to ensure that they add up to one for the treated and control units. This leads to the following estimator:

$$\hat{\tau}_{hir,2} = \sum_{i=1}^{N} Y_i \cdot \left(\frac{W_i}{\hat{e}(X_i)} - \frac{1-W_i}{1-\hat{e}(X_i)}\right) \Big/ \sum_{i=1}^{N} \left(\frac{W_i}{\hat{e}(X_i)} - \frac{1-W_i}{1-\hat{e}(X_i)}\right).$$

3.3 A New Estimator

The new estimator relies on estimating the unknown regression functions $\mu_1(x)$ and $\mu_0(x)$ through nonparametric regression of Y on X for the treatment and control subpopulations separately. These two nonparametric regressions are used to impute the counterfactual outcomes. For this reason we refer to the new estimator as the *imputation estimator*. The imputation estimator is a modification of Hahn's estimator. Like Hahn's estimator it estimates the average treatment effect as the average of the difference between the imputed outcomes

$$\hat{\tau}_{imp} = \frac{1}{N}\sum_{i=1}^{N} \left(\hat{\mu}_1(X_i) - \hat{\mu}_0(X_i)\right).$$

The difference with Hahn's estimator is that the two regression functions are estimated directly without first estimating the propensity score and the two conditional expectations $\mathbb{E}[Y \cdot W|X]$ and $\mathbb{E}[Y \cdot (1-W)|X]$. Hahns estimator also estimates $\mathbb{E}(Y_0|X)$ and $\mathbb{E}(Y_1|X)$, but by a different method. This implies that we only need to estimate two regression functions nonparametrically rather than three and this will simplify the optimal choice of smoothing parameters. Moreover the imputation estimator is linear in the nonparametrically estimated regression functions, and this further simplifies the population MSE of the average treatment effect and its estimator. Note that the imputation estimator still requires more unknown regression functions to be estimated than the HIR estimator which requires only estimation of the propensity score, but since the new estimator relies on estimation of different functions, it is not clear which is to be preferred in practice.

To provide additional insight into the structure of the problem, we introduce one final estimator, which combines features of the Hirano-Imbens-Ridder estimators and the imputation estimator:

$$\hat{\tau}_{mod} = \sum_{i=1}^{N} \left(\frac{W_i \cdot \hat{\mu}_1(X_i)}{\hat{e}(X_i)} - \frac{(1-W_i) \cdot \hat{\mu}_0(X_i)}{1-\hat{e}(X_i)} \right).$$

This estimator is a nonparametric version of the double-robust estimator proposed by Robins, Rotnitzky and Zhao (1995) and Rotnitzky and Robins (1995). It will play a role in the efficiency proof of the imputation estimator, but it may also be of independent interest.

The data consist of a random sample $(Y_i, W_i, X_i), i = 1, \ldots, N$. We can think of this as two random samples $(W_i Y_i(1), X_i), i = 1, \ldots, N$ and $((1-W_i)Y_i(0), X_i), i = 1, \ldots, N$. The first random sample is used to estimate $\mathbb{E}(Y(1))$ and the second to estimate $\mathbb{E}(Y(0))$. In both random samples the variable of interest is missing for some observations. The missing value is recorded as 0. Consider first the estimation of $\mathbb{E}(Y(1))$ using the random sample $(W_i Y_i(1), X_i), i = 1, \ldots, N$. (The estimation of $\mathbb{E}(Y(0))$ is completely analogous.) To simplify the notation we denote $Y_i(1)$ by Y_i. The conditional expectation $\mu(x) = \mathbb{E}(Y(1)|X) = \mathbb{E}(Y|X, W = 1)$ is estimated with the observations that have both Y_i and X_i, i.e. for which $W_i = 1$. The subsequent average is over the full sample, including the observations for which Y_i is not observed. The number of observations for which both Y_i and X_i are observed is $N_1 = \sum_{i=1}^{N} W_i$. In the current setup N_1 is proportional to $N \to \infty$, so that asymptotic bounds can be expressed as functions of N. Without loss of generality we arrange that data such that the first N_1 observations have $W_i = 1$.

It should be noted that there is another type of data that can be used to estimate $\mathbb{E}(Y(0)), \mathbb{E}(Y(1))$. In particular, we could have two independent random samples from the joint distributions of (Y, X) given $W = 1$ and $W = 0$ and in addition an independent random sample from the marginal population distribution of X. The imputation estimator can be computed with this information. Note that the propensity score cannot be identified unless we know $\Pr(D = 1)$ and without this information, neither the Hahn nor the weighting estimator can be used. Most of our results, and in particular results that compare the different estimators, are for a single sample. We will see that the order selection for the imputation estimator does not depend on the type of data that is used.

In order to implement these estimators we need estimators for the two regression functions and the propensity score. Following Newey (1995), we use series estimators for the two regression functions $\mu_w(x)$. Let K_w denote the number of basis functions in the series. As basis functions we use power series. Let $\lambda(d) = (\lambda_1, ..., \lambda_d)$ be a multi-index of dimension d, that is, a d-dimensional vector of non-negative integers, with $|\lambda(d)| = \sum_{k=1}^{d} \lambda_k$, and let $x^{\lambda(d)} = x_1^{\lambda_1} \ldots x_d^{\lambda_d}$. Consider a series $\{\lambda(r)\}_{r=1}^{\infty}$ containing all distinct vectors λ such that $|\lambda(r)|$ is nondecreasing. Let $p_r(x) = x^{\lambda(r)}$ and $P_r(x) = (p_1(x), ..., p_r(x))'$.

Under some conditions given below the matrix $\Omega_K = \mathbb{E}[P_K(X)P_K(X)'|W = 1]$ is non-

singular for all K (Newey, 1994). Hence we can construct a sequence of basis functions $R_K(x) = \Omega_K^{-1/2} P_K(x)$ with $\mathbb{E}[R_K(X)R_K(X)'] = I_K$. Let $R_{kK}(x)$ be the kth element of the vector $R_K(x)$. It will be convenient to work with this sequence of basis function $R_K(x)$. The vector of basis functions $R_K(x)$ is bounded by $\zeta(K) = \sup_{x\in\mathbb{X}} |R_K(x)|$. Newey (1994) shows that $\zeta(K) \leq CK$ if the basis functions are a power series.

The conditional expectation $\mu(x)$ is estimated by the series estimator $\hat{\mu}_K(x) = R_K'(x)\hat{\gamma}_K$ with $\hat{\gamma}_K$ the least squares estimator. The matrix $R_K'R_K$, with R_K the $N \times K$ matrix with ith row equal to $R_K(X_i)'$, may be singular, although Lemma A.3 in the appendix shows that this happens with probability going to zero. To deal with this case we define $\hat{\gamma}_K$ as

$$\hat{\gamma}_K = \begin{cases} 0 & \text{if } \lambda_{\min}(\hat{\Omega}_{K,N_1}) \leq 1/2, \\ \left(\sum_{i=1}^{N_1} R_K(X_i)R_K'(X_i)\right)^{-1} \sum_{i=1}^{N_1} R_K(X_i)Y_i & \text{otherwise,} \end{cases} \tag{3.3}$$

with $\hat{\Omega}_{K,N_1} = \frac{1}{N_1}\sum_{i=1}^{N_1} R_K(X_i)R_K(X_i)'$. Define 1_{N_1} to be an indicator for the event $\lambda_{\min}(\hat{\Omega}_{K,N_1}) \geq 1/2$. By Lemma A.3 in the appendix, it follows that a sufficient condition for $1_{N_1} \xrightarrow{p} 1$ is $\zeta(K)K^{1/2}N^{-1/2} \to 0$, i.e. the number of basis functions increases with N at a rate that ensures that $\hat{\Omega}_{K,N_1} \xrightarrow{p} I_K$, so that with a slight abuse of notation $1 - 1_{N_1} = O_p(\zeta(K)K^{1/2}N^{-1/2})$.

Given the estimated regression functions we estimate $\mathbb{E}(Y(1))$ as

$$\hat{\mu}_{Y(1)} = \frac{1}{N}\sum_{i=1}^{N} \hat{\mu}_K(X_i).$$

Note that the estimation of $\mathbb{E}(Y(0))$ is completely analogous.

The modified estimator $\hat{\tau}_{mod}$ also requires estimation of the propensity score. We use the series logit estimator (HIR) Let $L(z) = \exp(z)/(1+\exp(z))$ be the logistic cdf. The series logit estimator of the population propensity score $e(x)$ is $\hat{e}_L(x) = L(R_L(x)'\hat{\pi}_L)$, where

$$\hat{\pi}_L = \arg\max_{\pi} L_{N,L}(\pi), \tag{3.4}$$

for

$$L_{N,L}(\pi) = \sum_{i=1}^{N} \left(W_i \cdot \ln L(R_L(X_i)'\pi) + (1-W_i)\cdot \ln(1 - L(R_L(X_i)'\pi))\right). \tag{3.5}$$

3.4 First Order Equivalence of Estimators

We make the following assumptions.

Assumption 3.1 (Distribution of Covariates)
$X \in \mathbb{X} \subset \mathbb{R}^d$, where $\mathbb{X}$ is the Cartesian product of finite intervals $[x_{jL}, x_{jU}]$, $j = 1, \ldots, d$, with $-\infty < x_{jL} < x_{jU} < \infty$. The density of X is bounded away from zero on $\mathbb{X}$.

Assumption 3.2 (Propensity Score)

(*i*) *The propensity score is bounded away from zero and one on* $\mathbb{X}$.

(*ii*) *The propensity score is* s_e *times continuously differentiable on* $\mathbb{X}$.

Assumption 3.3 (Conditional Outcome Distributions)

(*i*) *The two regression functions* $\mu_w(x)$ *are* s_μ *times continuously differentiable on* $\mathbb{X}$.

(*ii*) *The conditional variance of* $Y_i(w)$ *given* $X_i = x$ *is bounded by* σ_w^2 *on* $\mathbb{X}$.

Because $e(x) = \Pr(W = 1|X = x)$ is bounded from 0 on $\mathbb{X}$, Assumption 3.1 implies that the density of the distribution of $X|W = 1$ is also bounded away from 0 on $\mathbb{X}$ (and that $\mathbb{X}$ is its support).

The properties of the estimators will follow from the following theorem:

Theorem 3.1 (Asymptotic Equivalence of $\hat{\tau}_{imp}$, $\hat{\tau}_{mod}$, $\hat{\tau}_{hir,1}$ and $\hat{\tau}_{hir,2}$)
Suppose assumptions 3.1-3.3 hold. Then
(*i*),

$$\sqrt{N}\cdot(\hat{\tau}_{imp} - \hat{\tau}_{mod}) = O_p\left(N^{-1/2}\zeta(K)^2\zeta(L)K^{1/2}L^{1/2}\right)+O_p\left(N^{-1/2}\zeta(K)^3K^{1/2}\right)+O_p\left(N^{-1/2}\zeta(L)^5L\right)$$

$$+O_p\left(N^{1/2}\zeta(K)^2\zeta(L)^2K^{-s_\mu/d}L^{-s_e/(2d)}\right)+O_p\left(\zeta(K)^2\zeta(L)L^{1/2}K^{-s_\mu/d}\right)+O_p\left(N^{1/2}\zeta(K)K^{-s_\mu/d}\right)$$
$$+O_p\left(\zeta(L)^4L^{1/2}L^{-s_e/(2d)}\right) + O_p\left(N^{1/2}\zeta(L)^2L^{-s_e/d}\right) + O_p\left(N^{1/2}\zeta(L)L^{-s_e/(2d)}\right)$$
$$+O_p\left(\zeta(K)^2\zeta(L)^2K^{1/2}L^{-s_e/(2d)}\right),$$

(*ii*),

$$\sqrt{N}\cdot(\hat{\tau}_{mod} - \hat{\tau}_{hir,1}) = O_p\left(N^{-1/2}\zeta(K)^2\zeta(L)K^{1/2}L^{1/2}\right) + O_p\left(N^{-1/2}\zeta(L)^5L\right)$$
$$+O_p\left(N^{1/2}\zeta(K)^2\zeta(L)^2K^{-s_\mu/d}L^{-s_e/(2d)}\right) + O_p\left(\zeta(K)\zeta(L)L^{1/2}K^{-s_\mu/d}\right)$$
$$+O_p\left(\zeta(L)^4L^{1/2}L^{-s_e/(2d)}\right) + O_p\left(\zeta(K)^2\zeta(L)^2K^{1/2}L^{-s_e/(2d)}\right)$$
$$+O_p\left(N^{1/2}\zeta(K)^2K^{1/2}K^{-s_e/d}\right) + O_p\left(N^{1/2}\zeta(K)K^{1/2}K^{-s_e/d-s_\mu/d}\right),$$

(*iii*),

$$\sqrt{N}\cdot(\hat{\tau}_{hir,1} - \hat{\tau}_{hir,2}) = O_p\left(N^{-1/2}\zeta(L)^5L\right) + O_p\left(\zeta(L)^4L^{1/2}L^{-s_e/(2d)}\right)$$
$$+O_p\left(N^{1/2}\zeta(L)L^{-s_e/(2d)}\right) + O_p\left(N^{1/2}\zeta(L)^2L^{-s_e/d}\right)$$

Proof: See appendix.

In the bounds it is implicitly assumed that $K_0 = K_1$. Alternatively, the bounds are the sum of the bounds in theorem 3.1 with K replaced by K_0 and K_1 respectively. To obtain bounds on the rate at which K and L increase with N, we need to choose a particular class of basis functions. For power series we have $\zeta(K) = K$. The next corollary gives the rates at which the estimators are asymptotically equivalent.

Corollary 3.1 *If $\zeta(K) = K$ and $K = N^{\nu_K}$, $L = N^{\nu_L}$, then $\hat{\tau}_{imp}$ and $\hat{\tau}_{mod}$ are asymptotically equivalent if $s_e/d > 9$ and*

$$\begin{aligned} \frac{1}{2\,(s_\mu/d-1)} &< \delta_K < \frac{1}{7} \\ \frac{1}{s_e/d-2} &< \delta_L < \frac{1}{12} \\ \frac{5}{s_e/d-4} &< \tfrac{\delta_L}{\delta_K} < \frac{2(s_\mu/d-4)}{3} \end{aligned}$$

Note that to satisfy the first two inequalities we need $s_\mu/d > 4$ and $s_e/d > 14$. With these restrictions the final inequality can be satisfied if $s_\mu/d > 5$ and in that case all inequalities can be satisfied simultaneously. Note that if $e(\cdot)$ is smoother, we can let L go to ∞ slower relative to K and the opposite is true if μ is smoother.

Corollary 3.2 *If $\zeta(K) = K$ and $K = N^{\nu_K}$, $L = N^{\nu_L}$, then $\hat{\tau}_{mod}$ and $\hat{\tau}_{hir,1}$ are asymptotically equivalent if $s_e/d > 9$ and*

$$\begin{aligned} & & \delta_L < & \frac{1}{12} \\ \frac{1}{2(s_e/d)-5} &< \delta_K < & \frac{1}{5} - \frac{3}{5}\delta_L \\ \frac{5}{s_e/d-4} &< \tfrac{\delta_L}{\delta_K} < & \frac{2(s_\mu/d-1)}{3} \end{aligned}$$

To satisfy the last inequality we need that $s_\mu/d > 3$. In that case we can satisfy all inequalities without further restrictions.

Corollary 3.3 *If $\zeta(K) = K$ and $K = N^{\nu_K}$, $L = N^{\nu_L}$, then $\hat{\tau}_{mod}$ and $\hat{\tau}_{hir,2}$ are asymptotically equivalent if $s_e/d > 9$ and*

$$\frac{1}{2(s_e/d-2)} < \delta_L < \frac{1}{12}$$

In the proof of part (iii) we require that the weighting estimator is consistent. HIR (2003) show that a sufficient condition is that

$$\frac{1}{4(s_e/d-1)} < \delta_L < \frac{1}{9}$$

If the inequality in corollary 3.2 holds, then this inequality is always satisfied.

4 A feasible MSE criterion

All three estimators contain a function or functions that are estimated nonparametrically. For the Hahn estimator we need nonparametric estimates of the propensity score and the conditional expectations of the product of the outcome and the treatment/control indicators given the covariates. For the HIR weighting estimator an estimator of the propensity score is needed, and for the imputation estimator introduced in section 3.3 we need estimators for the conditional means in the treatment and control populations. Both Hahn and HIR use series estimators for either the propensity score or the conditional expectations. That leaves the question how to select the order of the series. For a meaningful comparison of the performance of these asymptotically equivalent estimators such a selection rule is essential.

Despite its practical importance there has been little work on the selection of the nonparametric estimators. The only paper that we are aware of is Ichimura and Linton (2003) who consider bandwidth selection if the propensity score in the weighting estimator is estimated by a kernel nonparametric regression estimator. The current practice in propensity score matching, which is a nonparametric estimator that is different from the estimators considered in Section 3, is that the propensity score is selected using the balancing score property

$$W \perp X | e(X)$$

In practice this is implemented by stratifying the sample on the propensity score and testing whether the means of the covariates are the same for treatment and controls (see e.g. Dehejia and Wahba (2000)). This method of selecting the model for the propensity score focuses exclusively on the bias in the estimation of the treatment effect. This could lead to over-fitting of the propensity score and inflation of the variance of the estimator of the treatment effect.

We consider both the bias and the variance associated with a choice of the nonparametric function in the treatment effect estimator. As in Section 3we consider the estimation of $\mathbb{E}(Y(1))$ using the sample $(W_i Y_i(1), X_i), i = 1, \ldots, N$ and we denote $Y(1)$ by Y. The order selection for the estimation of $\mathbb{E}(Y(0))$ is completely analogous, with the understanding that the orders for the two parameters are chosen independently.

4.1 The MSE and its estimator

As in Li (1987) we consider a population in which the joint distribution of Y, X is such that

$$Y = \mu(X) + U$$

with $\mathbb{E}(U|X) = 0$ and $\mathrm{Var}(U|X) = \sigma^2$. Andrews (1991) has generalized Li's results to the heteroskedastic case and we could do the same. To concentrate on essentials first we maintain the assumption that U is homoskedastic. Note that by unconfoundedness $U \perp W|X$.

The data are as in Section 3.3, i.e. we have a sample $(W_i Y_i, X_i), i = 1, \ldots, N$ with as before $W \cdot Y = W \cdot Y(1)$. Without loss of generality we assume that Y is observed for the first

$N_1 = \sum_{i=1}^N W_i$ observations, and these observations are a random sample from the distribution of $Y, X|W = 1$. The imputation estimator for $\mu_Y = \mathbb{E}(Y)$ is

$$\hat{\tau}_{imp} = \hat{\mu}_{Y,K} = \frac{1}{N}\sum_{i=1}^N \hat{\mu}_{K,N_1}(X_i)$$

with

$$\hat{\mu}_{K,N_1}(x) = R_K(x)'(R_{K,N_1}'R_{K,N_1})^{-1}R_{K,N_1}'y$$

The subscript K is the number of basis functions used in the estimation of $\mu(x)$, and we use the notation

$$y_{N_1} = \begin{pmatrix} Y_1 \\ \vdots \\ Y_{N_1} \end{pmatrix} \qquad \mu_{N_1} = \begin{pmatrix} \mu(X_1) \\ \vdots \\ \mu(X_{N_1}) \end{pmatrix} \qquad u_{N_1} = \begin{pmatrix} U_1 \\ \vdots \\ U_{N_1} \end{pmatrix}$$

and

$$R_K(x) = \begin{pmatrix} R_{1K}(x) \\ \vdots \\ R_{KK}(x) \end{pmatrix} \quad R_{K,N_1} = \begin{pmatrix} R_K(X_1)' \\ \vdots \\ R_K(X_{N_1})' \end{pmatrix} \quad R_{K,N} = \begin{pmatrix} R_K(X_1)' \\ \vdots \\ R_K(X_N)' \end{pmatrix}$$

As is common in this literature we treat $X_i, i = 1, \ldots, N$ as constants. Alternatively, we can consider the derivation of the MSE as being conditional on $X_i, i = 1, \ldots, N$. In particular, when considering the population MSE it is convenient to average the MSE over X.

The MSE is obtained from

$$\sqrt{N}(\hat{\mu}_{Y,K} - \mu_Y) = \frac{1}{\sqrt{N}}\sum_{i=1}^N \left(\hat{\mu}_{K,N_1}(X_i) - \mathbb{E}_{Y|X,W=1}\left[\hat{\mu}_{K,N_1}(X_i)\right]\right) \tag{4.6}$$

$$+\frac{1}{\sqrt{N}}\sum_{i=1}^N \left(\mathbb{E}_{Y|X,W=1}\left[\hat{\mu}_{K,N_1}(X_i)\right] - \mu(X_i)\right) + \tag{4.7}$$

$$+\frac{1}{\sqrt{N}}\sum_{i=1}^N \left(\mu(X_i) - \mu_Y\right) \tag{4.8}$$

Because we treat $X_i, i = 1, \ldots, N$ as constants, we redefine the parameter of interest as

$$\mu_Y = \frac{1}{N}\sum_{i=1}^N \mu(X_i)$$

Hence the final term in the comparison, (4.8), can be omitted, and we need only consider the first two terms. This parameter is the sample treatment effect of Imbens (2004).

We now consider the first two terms separately. The first term, (4.6), corresponds to the variance and the second, (4.7), to the bias term in the MSE. The first term can be written as

$$V = \frac{1}{\sqrt{N}} \sum_{i=1}^{N} \left(\hat{\mu}_{K,N_1}(X_i) - \mathbb{E}_{Y|X,W=1} \left[\hat{\mu}_{K,N_1}(X_i) \right] \right) = \frac{1}{\sqrt{N}} \iota_N' R_{K,N} (R_{K,N_1}' R_{K,N_1})^{-1} R_{K,N_1}' u_{N_1}$$

so that the variance term of the MSE is

$$\mathbb{E}(V^2) = \sigma^2 \frac{1}{N} \iota_N' R_{K,N} (R_{K,N_1}' R_{K,N_1})^{-1} R_{K,N}' \iota_N = \frac{1}{N} \sigma^2 \left(\iota_N' R_{K,N} \right) \left(R_{K,N_1}' R_{K,N_1} \right)^{-1} \left(\iota_N' R_{K,N} \right)'$$

$$= \frac{1}{N} \sigma^2 \iota_N' \widetilde{M}_{K,N} \iota_N$$

with

$$\widetilde{M}_{K,N} = R_{K,N} \left(R_{K,N_1}' R_{K,N_1} \right)^{-1} R_{K,N}'$$

Because

$$\frac{R_{K,N}' \iota_N}{N} = \frac{1}{N} \sum_{i=1}^{N} r_K(X_i) \approx \frac{1}{N} \sum_{i=1}^{N} \frac{W_i}{e(X_i)} r_K(X_i) = \frac{1}{N} \sum_{i=1}^{N_1} \frac{1}{e(X_i)} r_K(X_i) = \frac{1}{N} R_{K,N_1}' a_{N_1}$$

($\approx$ means that left- and right-hand side have the same expectation and limit) with

$$a_{N_1} = \begin{pmatrix} \frac{1}{p(X_1)} \\ \vdots \\ \frac{1}{p(X_{N_1})} \end{pmatrix}$$

we obtain an alternative expression for the variance term

$$\mathbb{E}(V^2) = \frac{\sigma^2}{N} a_{N_1}' M_{K,N_1} a_{N_1}$$

with

$$M_{K,N_1} = R_{K,N_1} (R_{K,N_1}' R_{K,N_1})^{-1} R_{K,N_1}'$$

This expression is useful for the study of the relation between the smoothness of the propensity score and the variance.

The bias term is

$$B = \frac{1}{\sqrt{N}} \sum_{i=1}^{N} \left(\mathbb{E}_{Y|X,W=1} \left[\hat{\mu}_{K,N_1}(X_i) \right] - \mu(X_i) \right) = \frac{1}{\sqrt{N}} \sum_{i=1}^{N} \left(R_K(X_i)' (R_{K,N_1}' R_{K,N_1})^{-1} R_{K,N_1}' \mu_{N_1} - \mu(X_i) \right)$$

$$\approx \frac{1}{\sqrt{N}} \sum_{i=1}^{N} \frac{W_i}{e(X_i)} \left[R_K(X_i)' (R_{K,N_1}' R_{K,N_1})^{-1} R_{K,N_1}' \mu_{N_1} - \mu(X_i) \right]$$

$$= \frac{1}{\sqrt{N}} \sum_{i=1}^{N_1} \frac{1}{e(X_i)} \left[R_K(X_i)'(R'_{K,N_1} R_{K,N_1})^{-1} R'_{K,N_1} \mu_{N_1} - \mu(X_i) \right] = -\frac{1}{\sqrt{N}} a'_{N_1} A_{K,N_1} \mu_{N_1}$$

with $A_{K,N_1} = I - M_{K,N_1}$. Note that $\frac{a'_{N_1} A_{K,N_1} \mu_{N_1}}{N_1}$ is the covariance of the residuals of the regressions of μ_{N_1} on R_{K,N_1} and a_{N_1} on R_{K,N_1}, respectively. By Lemma A.5, A.6(ii), A.7 and A.8(ii) in the appendix, we have that

$$|a'_{N_1} A_{K,N_1} \mu_{N_1}| = O\left(N \zeta(K)^2 K^{1/2 - s_\mu/d - s_e/(2d)} \right)$$

so that the squared bias term is of order $O\left(N \zeta(K)^4 K^{1-2s_\mu/d - s_e/d}\right)$ with s_μ the number of continuous derivatives of $\mu(x)$, s_e the number of continuous derivatives of $e(x)$, and d the dimension of X.

Because $U \perp W|X$, we have that $\mathbb{E}(U|X, W = 1) = 0$, so that B and V are uncorrelated. Hence the population MSE is the sum of the variance of V^2 and B^2

$$E_N(K) = \frac{1}{N} \left(\sigma^2 \iota'_N \widetilde{M}_{K,N} \iota_N + (a'_{N_1} A_{K,N_1} \mu_{N_1})^2 \right) = \frac{1}{N} \left(\sigma^2 a'_{N_1} M_{K,N_1} a_{N_1} + (a'_{N_1} A_{K,N_1} \mu_{N_1})^2 \right) \tag{4.9}$$

The first expression on the right-hand side of (4.9) is the basis for an estimator of the MSE. There are several potential estimators of the bias term[2]. Let e_{K,N_1} be the vector of residuals of the regression of y_{N_1} on R_{K,N_1}, i.e. $e_{K,N_1} = A_{K,N_1} y_{N_1}$. Using

$$a'_{N_1} e_{K,N_1} = a'_{N_1} A_{K,N_1} \mu_{N_1} + a'_{N_1} A_{K,N_1} u_{N_1}$$

so that

$$(a'_{N_1} e_{K,N_1})^2 = (a'_{N_1} A_{K,N_1} \mu_{N_1})^2 + a'_{N_1} A_{K,N_1} u_{N_1} u'_{N_1} A_{K,N_1} a_{N_1} + 2 a'_{N_1} A_{K,N_1} \mu_{N_1} a'_{N_1} A_{K,N_1} u_{N_1}$$

Hence,

$$E\left[(a'_{N_1} e_{K,N_1})^2 \right] = (a'_{N_1} A_{K,N_1} \mu_{N_1})^2 + \sigma^2 a'_{N_1} A_{K,N_1} a_{N_1}.$$

This suggests the following estimator for the bias term:

$$(a'_{N_1} e_{K,N_1})^2 - \sigma^2 a'_{N_1} A_{K,N_1} a_{N_1}$$

Upon substitution of the estimate we obtain the estimated MSE

$$C_N(K) = \frac{1}{N} \left(\sigma^2 \iota'_N \widetilde{M}_{K,N} \iota_N + (a'_{N_1} e_{K,N_1})^2 - \sigma^2 a'_{N_1} A_{K,N_1} a_{N_1} \right) \tag{4.10}$$

[2]Note that $a' A_K \hat{\mu}_K = 0$, so that this obvious estimator cannot be used.

4.2 The population MSE

Before we consider properties of the estimated MSE (4.10) we study the behavior of the population MSE. Using the alternative expression for the variance term we find that if $N \to \infty$

$$\mathbb{E}(V^2) = \frac{\sigma^2}{N} a'_{N_1} M_{K,N_1} a_{N_1} \xrightarrow{p} \sigma^2 e \, \mathbb{E}\left[\frac{R_K(X)'}{e(X)}\middle| W=1\right] \Sigma_K^{-1} \mathbb{E}\left[\frac{R_K(X)}{e(X)}\middle| W=1\right]$$

with, because we use polynomials that are orthonormal with respsect to the distribution of $X|W=1$,

$$\Sigma_K = \mathbb{E}\left[R_K(X)R_K(X)'|W=1\right] = I_K$$

so that the variance term converges to

$$\sigma^2 e \sum_{k=1}^{K} \left(\mathbb{E}\left[\frac{R_{kK}(X)}{e(X)}\middle| W=1\right]\right)^2$$

Because

$$\frac{1}{N_1} a'_{N_1} A_{K,N_1} \mu_{N_1} \xrightarrow{p} \mathbb{E}\left[\frac{\mu(X)}{e(X)}\middle| W=1\right] - \sum_{k=1}^{K} \mathbb{E}\left[\frac{R_{kK}(X)}{e(X)}\middle| W=1\right] \mathbb{E}\left[R_{kK}(X)\mu(X)|\, W=1\right]$$

the bias term B is asymptotically equal to (ratio converges to 1)

$$\sqrt{N} e \left(\mathbb{E}\left[\frac{\mu(X)}{e(X)}\middle| W=1\right] - \sum_{k=1}^{K} \mathbb{E}\left[\frac{R_{kK}(X)}{e(X)}\middle| W=1\right] \mathbb{E}\left[R_{kK}(X)\mu(X)|\, W=1\right]\right)$$

To obtain simple expressions for the bias and variance terms, we assume that

$$e(x) = \frac{1}{\gamma'_{K_e} R_{K_e}(x)} \qquad \mu(x) = \delta'_{K_\mu} R_{K_\mu}(x) \tag{4.11}$$

Because $R_{1K}(x) \equiv 1$ and the support of X is bounded, we can ensure that $e(x)$ is a proper probability on the support of X. We have for this choice

$$\begin{aligned} \mathbb{E}\left[\frac{R_{kK}(X)}{e(X)}\middle| W=1\right] &= \gamma_k \qquad \text{if } k \le K_e \\ &= 0 \qquad \text{if } k > K_e \end{aligned}$$

and analogously for $\mu(x)$, so that

$$\mathbb{E}\left[\frac{\mu(X)}{e(X)}\middle| W=1\right] = \sum_{k=1}^{\min\{K_e,K_\mu\}} \gamma_k \delta_k$$

We conclude that the population MSE is approximately

$$E_N(K) = \sigma^2 e \sum_{k=1}^{\min\{K,K_e\}} \gamma_k^2 + N e^2 \left(\sum_{k=K+1}^{\min\{K_e,K_\mu\}} \gamma_k \delta_k\right)^2$$

For fixed N the variance term increases with K. The squared bias decreases with K if γ_k and δ_k have the same sign. The bias term need not be monotonic in K. If $K_e < K_\mu$, then the variance term is constant and the squared bias is 0 if $K \geq K_e$. A special case is random assignment of treatment with $K_e = 1$. In this case $E_N(K)$ is constant, suggesting the choice $K = 1$ that in this case is clearly optimal. Because for random assignment $a_{N_1} = \frac{1}{\Pr(W=1)}\iota_{N1}$ both the squared bias and its estimate are 0, so that the estimated MSE is minimized for $K = 1$ irrespective of the sample size.

If (4.11) holds then the squared bias term in the population MSE in (4.9) is 0 for $K \geq \min(K_e, K_\mu)$. This is true irrespective of the sample size. This implies that the population MSE is minimized at $K = \min(K_e, K_\mu)$. The estimate of the squared bias term in (4.10) is 0 if $K \geq K_e$, but in large samples $a'_{N_1} e_{K,N_1}/N_1$ is approximately 0 if $K_\mu \leq K < K_e$ (see section 4.3) so that in large samples the estimated MSE is also minimized at $K = \min(K_e, K_\mu)$.

If we consider the number of basis functions that are needed to approximate either $\mu(x)$ or $1/e(x)$ as an index of smoothness, we conclude that the relative smoothness of these functions determines the number of basis functions that minimizes the population MSE. If the (the inverse of) propensity score is smoother than the conditional mean, we may need only a small number of basis functions, even if the conditional mean is unsmooth. This can be seen as oversmoothing of the conditional mean. Often in two-step semiparametric procedures the first step nonparametric estimator is undersmoothed to deal with bias in the second stage parametric estimator. If the weights are smooth the second stage bias will be small and we can safely use a smooth first-stage nonparametric estimator. If $K_e < K_\mu$ the population MSE is constant for $K \geq K_e$ with K_e the minimizer of the MSE. In that case it seems natural to select the smallest order that minimizes the population MSE. The fact that the inverse of the propensity score plays an important role seems surprising, but is less so, if one considers that this inverse reweights the sample in which Y is observed to the population.

It is instructive to compare the population MSE with Li's (1986) average squared error criterion. This criterion is used to select the estimator that approximates $\mu(x)$ well. As we noted above, it may not be optimal to select an estimator that estimates the conditional expectation well. In particular, if the propensity score is smoother than the conditional expectation, Li's criterion may suggest a number of basis functions that is too large. To obtain intuitive results we assume again (4.11). Li's average squared error criterion is

$$L_N(K) = \frac{1}{N_1}\sum_{i=1}^{N_1}(\hat{\mu}_{K,N_1}(X_i) - \mu(X_i))^2$$

i.e. it is average squared deviation in the sample in which we observe both Y and X. The expected value of the variance term is

$$\sigma^2 \frac{1}{N_1}\sum_{i=1}^{N_1} R_K(X_i)'(R'_K R_K)^{-1} R_K(X_i) = \sigma^2 \frac{K}{N_1}$$

The bias term is equal to

$$\frac{1}{N_1}\mu'_{N_1}\mu_{N_1} - \frac{1}{N_1}\mu' R_{K,N_1}(R'_{K,N_1}R_{K,N_1})^{-1}R'_{K,n_1}\mu_{N_1}$$

which converges to $\sum_{k=K+1}^{K_0}\delta_k^2$. Hence Li's criterion is

$$\mathbb{E}\left[L_N(K)\right] = \sigma^2\frac{K}{N_1} + \sum_{k=K+1}^{K_\mu}\delta_k^2 \tag{4.12}$$

which we compare with

$$E_N(K) = e\left(\sigma^2 \sum_{k=1}^{\min\{K,K_e\}}\gamma_K^2 + N_1\left(\sum_{k=K+1}^{\min\{K_e,K_\mu\}}\gamma_k\delta_k\right)^2\right)$$

These two criteria are clearly different. For instance, for fixed N, $\mathbb{E}\left[L_N(K+1)\right] \leq \mathbb{E}\left[L_N(K)\right]$ if and only if $t_{K+1}^2 \geq 1$ with

$$t_{K+1} = \frac{\delta_{K+1}}{\frac{\sigma}{\sqrt{N}}}$$

the asymptotic t-ratio for δ_{K+1}. If we take $\gamma_k = \gamma$ and $K_e = K_\mu$, then for our criterion we find that it decreases if and only if

$$t_{K+1} \geq \sqrt{1 + T_{K+2}^2} - T_{K+2}$$

or

$$t_{K+1} \leq -\sqrt{1 + T_{K+2}^2} - T_{K+2}$$

with $T_{K+2} = \sum_{k=K+2}^{K_\mu} t_k$.

4.3 Optimality of the minimizer of the estimated MSE

Let $\mathcal{K}_N = \{k = 1, 2, \ldots | N^{\nu_0} \leq k \leq N^{\nu_1}\}$ be the set of positive integers between N^{ν_0} and N^{ν_1}. We will need $\nu_0 > 0$. The upper bound can be $\nu_1 = 1$. Define $\hat{K}$ and K^* as

$$\hat{K}_N = \text{argmin}_{K\in\mathcal{K}_N} C_N(K), \qquad \text{and } K_N^* = \text{argmin}_{K\in\mathcal{K}_N} E_N(K).$$

We will show that $\hat{K}$ is optimal in the sense that (Li, 1987)

$$\frac{E_N(\hat{K})}{\inf_{K\in\mathcal{K}_N} E_N(K)} \xrightarrow{p} 1 \tag{4.13}$$

This does not imply that the difference between $\hat{K}$ and K^* converges to 0.

We make the following assumptions

Assumption 4.1 $\mathbb{E}\left[U^{2m}\right] < \infty$ *for some integer* m*.*

Note that this implies that

$$\mathbb{E}\left[U^{2m}|W=1\right] < \infty$$

which is what is used in the proof.

Assumption 4.2 *For the* m *satisfying Assumption 4.1,*

$$N^{-\nu_1+\nu_0 m(s_e/(2d)-1)} \inf_{K\in\mathcal{K}_N} E_N(K)^{\frac{m}{2}} \to \infty$$

with d *the dimension of* X*.*

Because the population MSE $E_N(K)$ is bounded from below by the positive variance term, a sufficient condition for assumption 4.2 is that $s_e > 2d\frac{m+\nu_1/\nu_0}{m}$. For sufficiently large m this holds under the rate assumptions in the Corolaries 3.1-3.3.

The sufficient condition illustrates an important issue with the MSE calculations. The optimal value K_N^* needs to increase fast enough so that $K_N^* > N^{\nu_0}$. This rules out particularly smooth functions the same way Li's approach rules out regression functions that are exact polynomials for with a finite number of nonzero coefficients.

Theorem 4.1 *If assumptions 4.1-4.2 hold, then*

$$\frac{E_N(\hat{K})}{\inf_{K\in\mathcal{K}_N} E_N(K)} \xrightarrow{p} 1$$

Proof See appendix.

4.4 Simulation results

The finite sample performance of our estimator is investigated in a small number of sampling experiments. The sampling experiments are limited and do not address a number of important questions, as the choice of basis functions, the estimation of σ^2, the order of basis functions, in particular when there are many covariates etc. The main issues that we investigate are the finite sample performance of the estimated MSE, and the behavior of the population MSE, in particular the role of the smoothness of the $\mu(x)$ and $e(x)$.

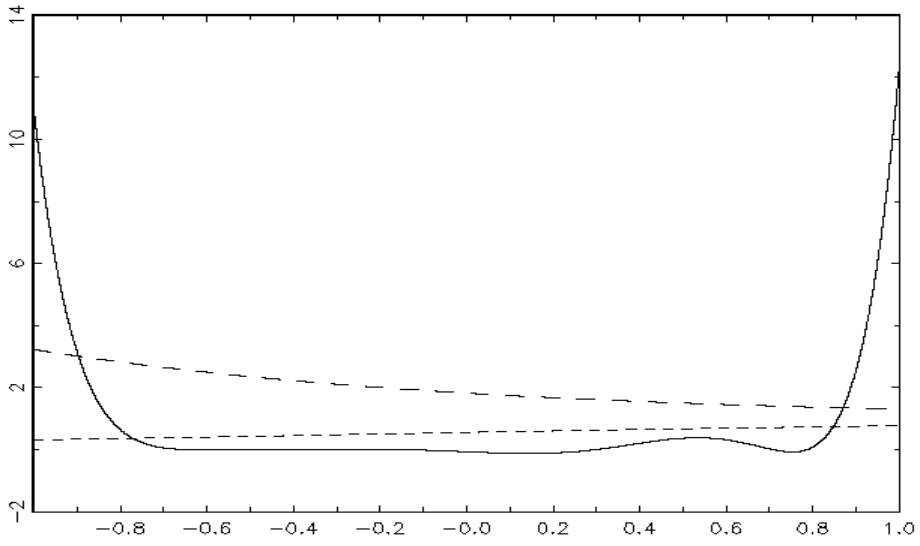
The population model is

$$Y = \mu(X) + U$$

with

$$\mu(x) = (x+.6)(x+.42)(x+.3)(x+.2)(x-.1)(x-.3)(x-.4)(x-.5)+$$

Figure 1: Conditional mean function, propensity score and inverse propensity score



$$+50(x+.65)(x+.55)(x+.3)*(x+.2)(x+.4)(x-.3)(x-.8)(x-.7)$$

X is a scalar variable and U is normal with mean 0 and variance 5. We choose a simple logit model for $e(x)$

$$e(x) = \frac{e^{.2+1x}}{1+e^{.2+1x}}$$

Graphs of the functions $\mu(x)$, $e(z)$ and $1/e(x)$ are given in figure 1.

As basis functions we choose the Legendre polynomials. These polynomials are defined on $[-1,1]$ and are orthogonal with weight function equal to 1 on this interval. For that reason we choose the marginal distribution of X such that $X|W=1$ is a uniform distribution on $[-1,1]$, i.e. the distribution of X has density

$$f(x) = \frac{\Pr(W=1)}{2}\frac{1}{p(x)} \qquad -1 \leq x \leq 1$$

Of course, in practice it may not be feasible to choose polynomials that are orthogonal with a weight function that is equal to the distribution $X|W=1$. The results in sections 4.1 and 4.3 do not depend on this choice, and the only reason for setting the simulation up this way is to check the results in section 4.2

Table 1 gives some statistics for the data generating process. The fraction in the population with an observed Y is .5089. The mean of Y is smaller in the subpopulation with observed Y than in the full population. The propensity score is between .31 and .77, i.e. the deviation from random assignment is moderate and the weights are not too large (see also figure 1). The R^2 is that of the population model.

Table 1: Statistics sampling experiment

	Mean	Std dev
Y	.9146	3.1248
$Y\|W=1$	.8021	3.0511
W	.5089	
R^2	.4878	

Table 2: Population and estimated MSE

	mean $\tilde{K}$	mean $\hat{K}$	min $\hat{K}$	max $\hat{K}$	frac $\hat{K}=\tilde{K}$	$\frac{E_N(\hat{K})}{E_N(\tilde{K})}$
$N=100$	2.4	2.1	1	5	.669	1.0487
$N=1000$	3.6	3.4	1	5	.773	1.0286
$N=5000$	4.9	4.8	3	6	.923	1.0001

We consider three sample sizes $N = 100, 1000, 5000$. The number of repetitions is 1000. The results are in table 2 and the figures.

The final column is in line with theorem 4.1. That theorem does not imply that the minimizer of the population and estimated MSE are closer if the sample size increases. However, this is what we find in this experiment. Indeed if we consider even larger sample sizes (not reported), we find that both the population and estimated MSE are minimal for $K = 5$. Hence, in this example the weights are sufficiently smooth (but not exactly a linear combination of basis functions) that the bias is essentially 0 and the variance constant if the number of basis functions is smaller than the order of the polynomial for $\mu(x)$.

In the figures 1-3 we report the population MSE and its estimator and the poulation squared bias and variance terms. The first thing to note is that the estimator of the population MSE is acurate, in particular for larger sample sizes. The estimate has to be accurate for large K because in that case the bias is very small compared to the variance, and the variance term need not be estimated. For sample size 100 the variance seems to increase with K this is a small sample phenomenon, because the variability of the variance term for larger K is big. For larger sample sizes the behavior of the population MSE is as if the inverse propensity score is a polynomial of order 5. It should be noted that for if the sample size is small, the population MSE may have local minima. In particular, it occurs that the population MSE is minimal for

$K = 1$ increases if $K = 2$ and then decreases again for $K \geq 2$ in some cases to a value that is larger than the minimum at $K = 1$. In this case the preferred estimate is the biased mean in the subsample in which Y is observed. As predicted by theorem 4.1 local minima disappear if the sample size increases.

4.5 Empirical application

In the empirical application we use data from the Great Avenue to INdependence (GAIN) experiments. These experimental evaluations of job training and job search assistance programs took place in the 1990's in a number of locations in California. As the evaluations are based on formal randomization within the sites, evaluations of the within-site effects are straightforward. Here we focus on comparisons of the control groups in different sites.

Such comparisons can be useful to assess whether for the program evaluations it is important to have randomizations. Specifically, to analyze the question whether controls in one location (e.g., Los Angeles) could have provided a comparison group for the program in a second location (e.g., Riverside), one could investigate whether conditional on covariates there is a systematic difference (e.g., a non-zero average difference) between the two groups. This amounts to estimating an average treatment effect on the control group data from the two locations with the location indicator as the treatment dummy.

We compare the control groups in Riverside and Los Angeles with 313 and 1022 observations in the control groups respectively. For each individual we have individual background characteristics, including age, ethnicity (hispanic, black, or other), an indicator for high school graduation, an indicator for the presence of exactly one child (all individuals have at least one child), and an indicator for the presence of children under the age of 5, and ten quarters of earnings data. As outcome we use earnings in the year after the program. Table 3 presents summary statistics for the covariates by site. All earnings data are in thousands of dollars.

The table suggests that the two locations have some differences, especially in the average earnings prior to the enrollment in the program, but only limited differences in individual background charateristics other than ethnicity (the control group in LA has many more Blacks relative to Whites compared to Riverside, with similar proportions of Hispanics). The full population consist of the control group population in either Los Angeles or Riverside. The population means $\mathbb{E}(Y(0))$ and $\mathbb{E}(Y(1))$ are the mean earnings in the case that the combined control population would live in either Los Angeles or Riverside. We estimate the mean for Los Angeles county.

The sequence of models we consider is indexed by the number of pre-program earnings periods we include in the regression function. All models include the six individual background characteristics plus age-squared. The ten models then differ by the number of pre-program quarters of earnings data they include. The result is in figure 5 where the top graph is for an estimate of Li's criterium, Mallow's C_p and the bottom graph is for our estimated MSE. Note

Figure 2: Population and estimated MSE; population squared bias and variance: $N = 100$

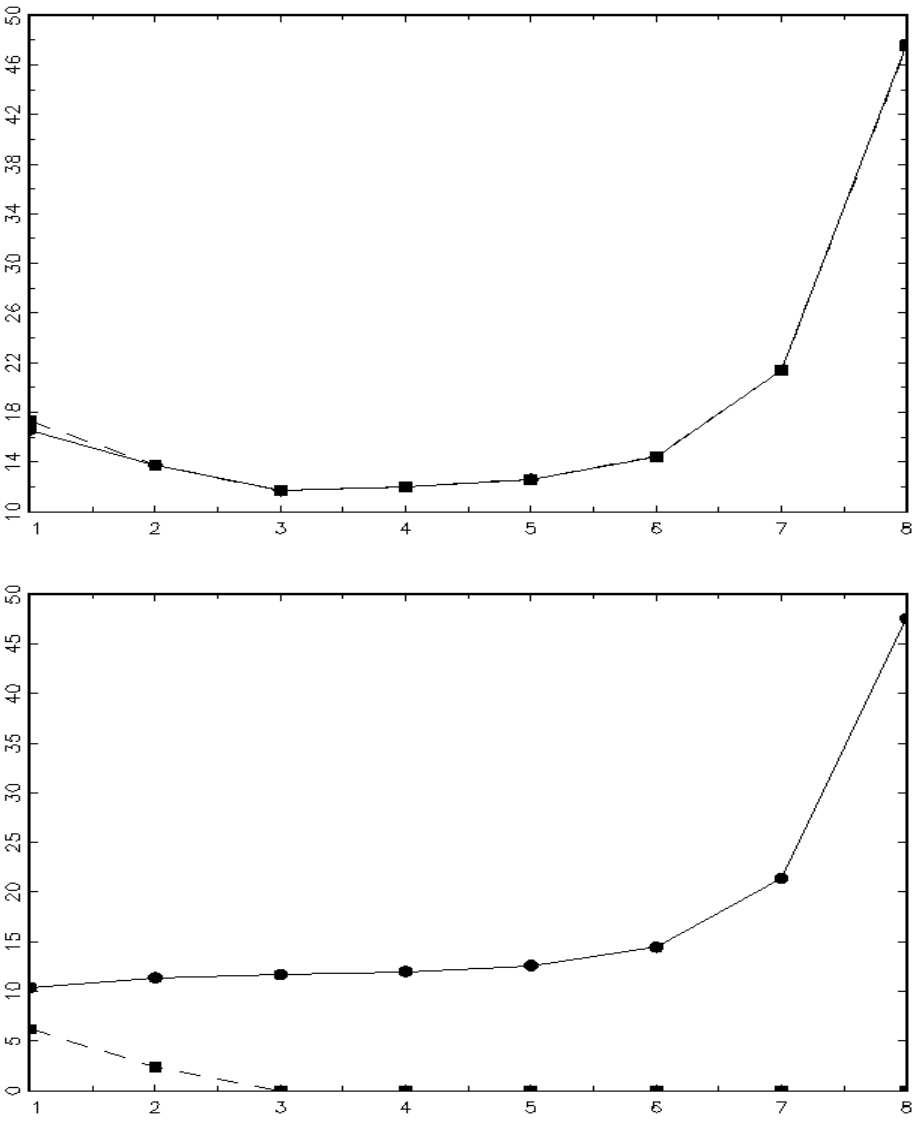


Figure 3: Population and estimated MSE; population squared bias and variance: $N = 1000$

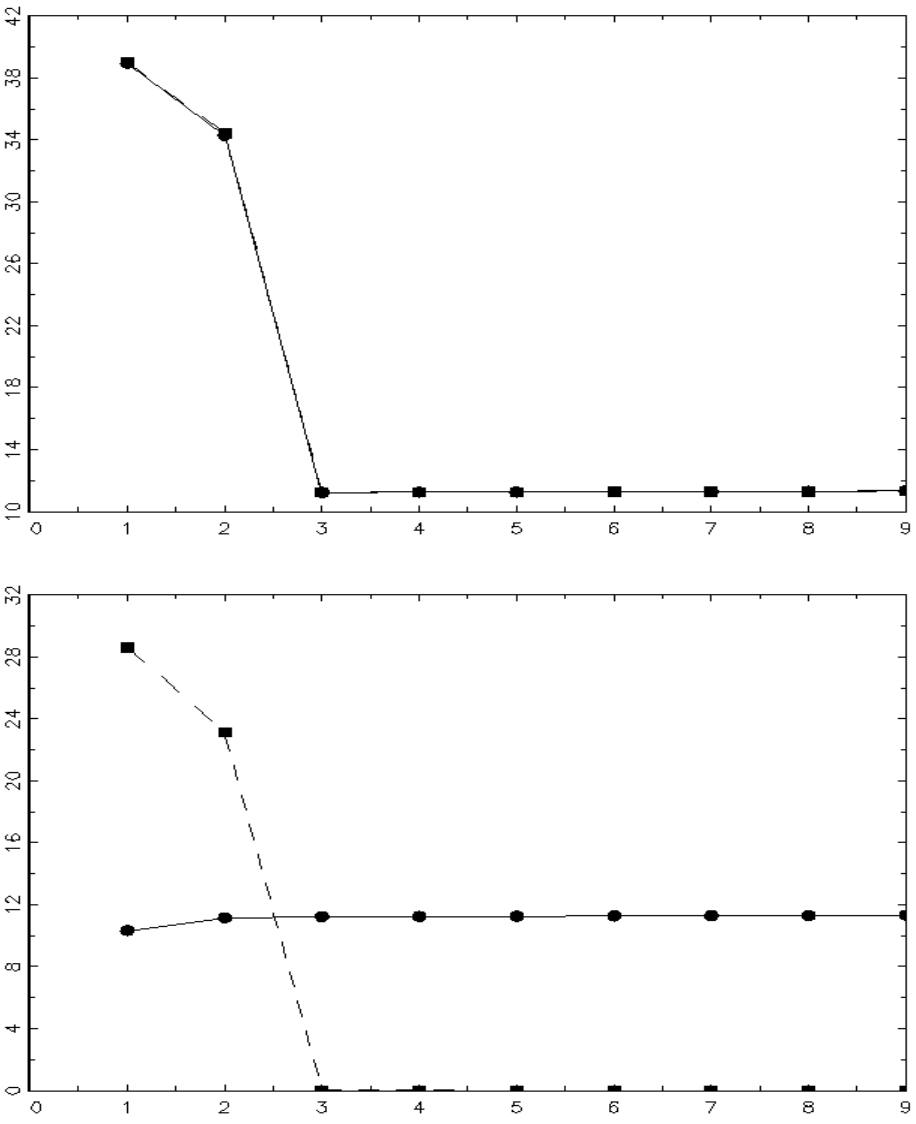


Figure 4: Population and estimated MSE; population squared bias and variance: $N = 5000$

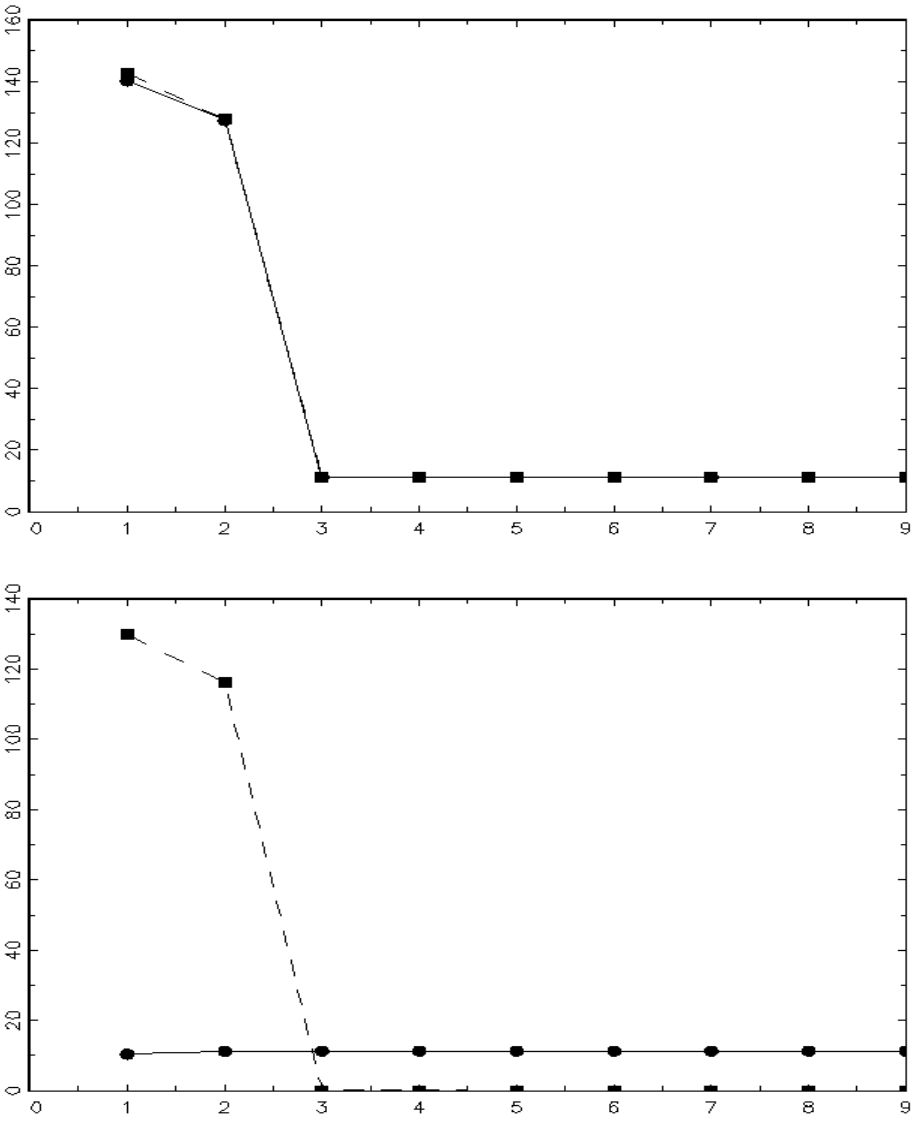


Table 3: GAIN Data: Summary Statistics

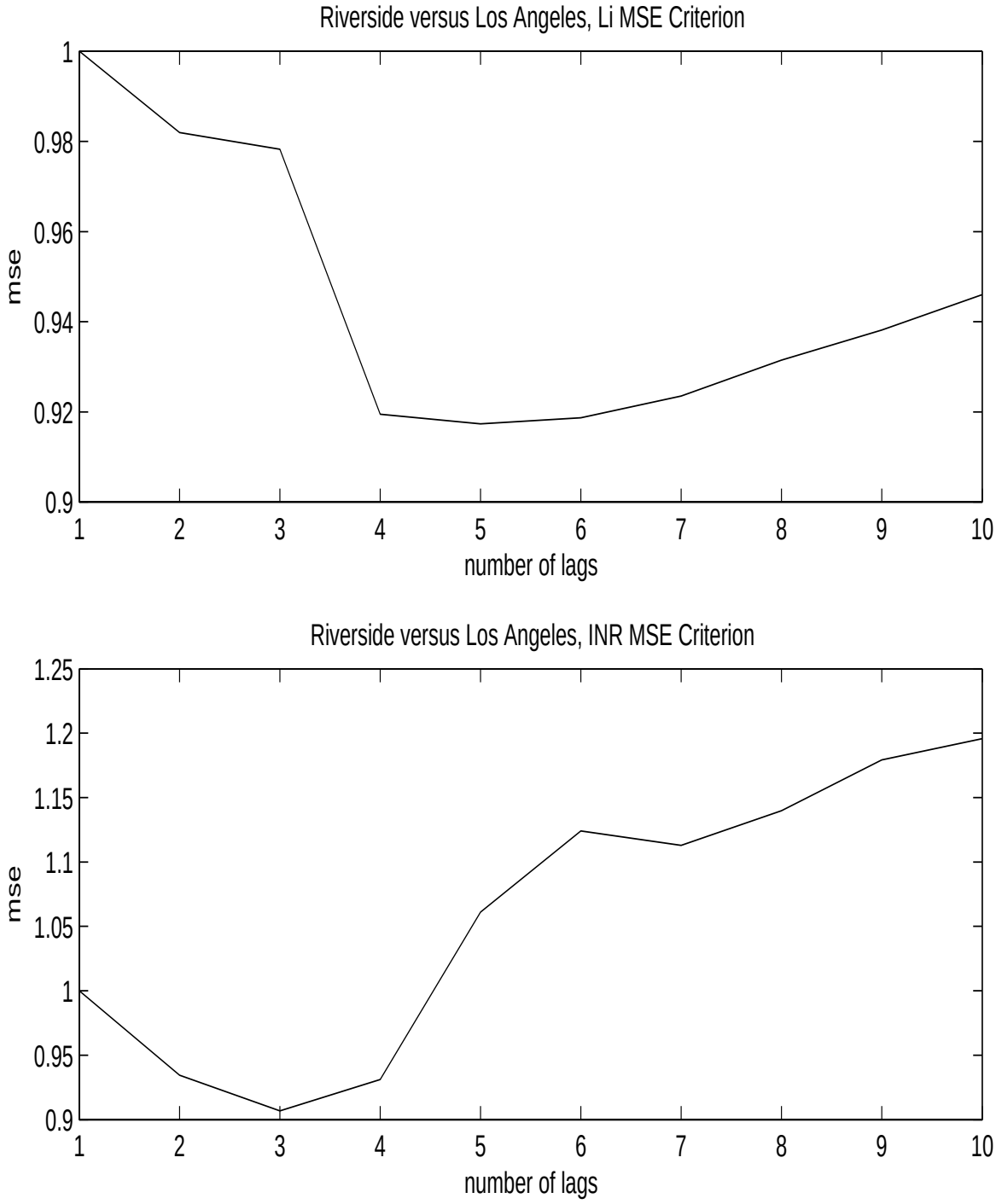
	All				Riverside (313)		Los Angeles (1022)	
	mean	s.d.	min	max	mean	s.d.	mean	s.d.
Age	37.7	8.6	16	66	34.4	8.7	38.7	8.3
Hispanic	0.36	0.48	0	1	0.37	0.48	0.36	0.48
Black	0.37	0.48	0	1	0.17	0.37	0.43	0.50
HS Dipl	0.20	0.40	0	1	0.21	0.41	0.19	0.39
1 Child	0.32	0.47	0	1	0.35	0.48	0.31	0.46
Child [0,5]	0.12	0.32	0	1	0.13	0.34	0.11	0.31
Earnings Quarter -1	0.20	0.81	0	9.09	0.16	0.64	0.21	0.86
Earnings Quarter -2	0.20	0.79	0	9.73	0.19	0.71	0.20	0.82
Earnings Quarter -3	0.20	0.77	0	9.37	0.13	0.51	0.22	0.83
Earnings Quarter -4	0.20	0.81	0	11.10	0.14	0.49	0.22	0.89
Earnings Quarter -5	0.20	0.81	0	11.53	0.17	0.68	0.21	0.85
Earnings Quarter -6	0.19	0.76	0	9.62	0.17	0.67	0.19	0.79
Earnings Quarter -7	0.18	0.71	0	9.67	0.17	0.61	0.18	0.74
Earnings Quarter -8	0.16	0.71	0	10.32	0.19	0.66	0.16	0.72
Earnings Quarter -9	0.18	0.73	0	8.46	0.26	0.87	0.16	0.69
Earnings Quarter -10	0.18	0.76	0	10.39	0.27	0.88	0.16	0.71
Earnings Year 1	1.13	3.30	0	37.28	0.86	2.40	1.21	3.53

that C_p is minimal for $K = 5$ while our estimated MSE suggests that K should be chosen as 3.

5 Conclusion

Although the method that we propose to select the non-parametric first component in a two-step estimator of the average treatment effect works well in a simulation study and an empirical example, much work has to be done before it will be a fully automatic feature of the estimation procedure. First, we need more insight in the performance of our procedure in a wider array of settings. In particular, the choice of basis functions in the case that the covariate vector has more than one variable is important. Second, the choice of the variance parameter σ^2 may be important, at least in finite samples. The current practice is to use the variance estimate in a 'large' regression model.

Figure 5: Li and INR MSE criterium



The order selection is particularly simple for the imputation estimator that is linear in the nonparametric estimator. One would expect that the qualitative results on the relation between the smoothness of the regression function, the propensity score and the population MSE holds also for estimators that are nonlinear in the nonparametric estimators, as the HIR weighting estimator.

A Appendix

For matrices A we use the norm $|A| = (\text{tr}(A'A))^{1/2}$.[3] For reference we state some of the properities of this norm.

Lemma A.1 (Properties of Norm) *For conformable matrices A and B,*
(i), $|AB| \leq |A| \cdot |B|$,
(ii), $|A'BA| \leq |B| \cdot |A'A|$.
If B is positive semi-definite and symmetric, then, for $\lambda_{\max}(B)$ equal to the maximum eigenvalue of B,
(iii), $\text{tr}(A'BA) \leq |A|^2 \cdot \lambda_{\max}(B)$,
(iv), $|AB| \leq |A| \cdot \lambda_{\max}(B)$,
(v), $|BA| \leq |A| \cdot \lambda_{\max}(B)$.

Proof: Omitted.

The data consist of a random sample $(Y_i, W_i, X_i), i = 1, \ldots, N$. We can think of this as two random samples $(W_iY_i(1), X_i), i = 1, \ldots, N$ and $((1 - W_i)Y_i(0), X_i), i = 1, \ldots, N$. The first random sample is used to estimate $\mathbb{E}(Y(1))$ and the second to estimate $\mathbb{E}(Y(0))$. In both random samples the variable of interest is missing for some observations. The missing value is 0. In this appendix we only consider the estimation of $\mathbb{E}(Y(1))$ using the random sample $(W_iY_i(1), X_i), i = 1, \ldots, N$. The estimation of $\mathbb{E}(Y(0))$ is completely analogous. To simplify the notation we denote $Y(1)$ by Y. The conditional expectation $E(Y|X = x) \equiv \mu(x)$ is estimated with the observations that have both Y_i and X_i, i.e. for which $W_i = 1$. The subsequent average is over the full sample, including the observations for which Y is not observed. If it does not cause confusion, we suppress the condition $W = 1$ in moments of the (joint) distribution(s) of Y, X for the subpopulation in which both Y and X are observed. This makes no difference for the conditional distribution of Y given $X = x$ that by assumption does not depend on W. The number of observations for which both Y and X are observed is $N_1 = \sum_{i=1}^{N} W_i$. In the current setup N_1 is proportional to $N \to \infty$, so that asymptotic bounds are expressed as functions of N. Without loss of generality we arrange that data such that the first N_1 observations have $W_i = 1$.

We use a series estimator for the regression function $\mu(x)$. Let K denote the number of terms in the series. As the basis functions we use power series. Let $\lambda(d) = (\lambda_1, ..., \lambda_d)$ be a multi-index of dimension d, that is, a d-dimensional vector of non-negative integers, with $|\lambda(d)| = \sum_{k=1}^{d} \lambda_k$, and let $x^{\lambda(d)} = x_1^{\lambda_1} \ldots x_d^{\lambda_d}$. Consider a series $\{\lambda(r)\}_{r=1}^{\infty}$ containing all distinct vectors such that $|\lambda(r)|$ is nondecreasing. Let $p_r(x) = x^{\lambda(r)}$, where $P_r(x) = (p_1(x), ..., p_r(x))'$.

Assumption A.1 *$X \in \mathbb{X}$, where $\mathbb{X}$ is the Cartesian product of intervals $[x_{jL}, x_{jU}]$, $j = 1, \ldots, d$, with $x_{jL} < x_{jU}$. The density of X is bounded away from zero on $\mathbb{X}$.*

If we assume, as we do in assumption A.5, that $e(x) = \Pr(W = 1|X = x)$ is bounded from 0 on $\mathbb{X}$, assumption A.1 implies that the density of the distribution of $X|W = 1$ is also bounded away from 0 on $\mathbb{X}$ (and that $\mathbb{X}$ is its support). Given assumption A.1 the matrix $\Omega_K = \mathbb{E}[P_K(X)P_K(X)']$ is nonsingular for all K (Newey, 1994)[4]. Hence we can construct a sequence of basis functions $R_K(x) = \Omega_K^{-1/2} P_K(x)$

[3] Because this norm coincides with the usual Euclidean norm if A is a vector or a scalar, we use the same notation for the scalar, vector or matrix case. The properties of the norm depend in part on whether A is a scalar, vector or matrix.

[4] By the convention adopted above we omit the condition $W = 1$ in the (conditional) expectation.

with $\mathbb{E}[R_K(X)R_K(X)'] = I_K$. Let $R_{kK}(x)$ be the kth element of the vector $R_K(x)$. It will be convenient to work with this sequence of basis function $R_K(x)$. Define

$$\zeta(K) = \sup_{x\in\mathbb{X}} |R_K(x)|.$$

Lemma A.2 (Newey, 1994)

$$\zeta(K) = O(K).$$

Let R_K be the $N_1 \times K$ matrix with ith row equal to $R_K'(X_i)$, and let $\hat{\Omega}_{K,N_1} = R_K'R_K/N_1$.

Lemma A.3 (Newey, 1995)

$$|\hat{\Omega}_{K,N_1} - I_K| = O_p\left(\zeta(K)K^{1/2}N^{-1/2}\right)$$

Proof: By Cauchy-Schwartz

$$\mathbb{E}[|\hat{\Omega}_{K,N_1} - I_K|] \leq \sqrt{\mathbb{E}[|\hat{\Omega}_{K,N_1} - I_K|^2]} = \sqrt{\frac{1}{N^2}\mathbb{E}\left[\text{tr}\left(\left(\sum_{i=1}^N (R_K(X_i)R_K(X_i)' - I_K)\right)^2\right)\right]} =$$

$$\frac{1}{N}\sqrt{\text{tr}\left(\sum_{i=1}^N \mathbb{E}\left[(R_K(X_i)R_K(X_i)' - I_K)^2\right]\right)} \leq \frac{1}{N}\sqrt{\sum_{i=1}^N \mathbb{E}\left[\text{tr}(R_K(X_i)R_K(X_i)'R_K(X_i)R_K(X_i)')\right]} =$$

$$\frac{1}{\sqrt{N}}\sqrt{\mathbb{E}\left[(R_K(X_i)'R_K(X_i))^2\right]} \leq \frac{\sqrt{\sup_{x\in\mathbb{X}} R_K(x)'R_K(x)}}{\sqrt{N}}\sqrt{\mathbb{E}\left[R_K(X_i)'R_K(X_i)\right]} = O\left(\zeta(K)K^{1/2}N^{-1/2}\right)$$

The result follows by the Markov inequality. Note that this rate is faster than that in Newey (1995), because we take the basis functions as orthonormal. □

Let $U_i = Y_i - \mu(X_i)$. Let $\mathbf{U}$, $\mathbf{Y}$, and $\mathbf{X}$ be the N_1 vectors and $N_1 \times d$ matrix with ith row equal to U_i, Y_i, and X_i'. Let 1_{N_1} be the indicator for the event $\lambda_{\min}(\hat{\Omega}_{K,N_1}) > 1/2$.

Assumption A.2

$$\sup_{x\in\mathbb{X}} \text{Var}(Y|X) \leq \bar{\sigma}^2 < \infty.$$

Lemma A.4 *(i),*

$$1_{N_1} \cdot \left|\hat{\Omega}_{K,N_1}^{-1/2} R_K'\mathbf{U}/N_1\right| = O_p(K^{1/2}N^{-1/2}),$$

and (ii),

$$1_{N_1} \cdot \left|\hat{\Omega}_{K,N_1}^{-1} R_K'\mathbf{U}/N_1\right| = O_p(K^{1/2}N^{-1/2}),$$

Proof: First we prove (i).

$$\mathbb{E}\left[1_{N_1} \cdot \left|\hat{\Omega}_{K,N_1}^{-1/2} R_K'\mathbf{U}/N_1\right|^2 \,\middle|\, X\right]$$

$$= \mathbb{E}\left[1_{N_1} \cdot \mathbf{U}'R_K\hat{\Omega}_{K,N_1}^{-1} R_K'U/N_1^2 \middle| X\right]$$

$$
\begin{aligned}
&= \mathbb{E}\left[1_{N_1} \cdot \mathbf{U}' R_K (R_K' R_K)^{-1} R_K' \mathbf{U}/N_1 \middle| X\right] \\
&= \mathbb{E}\left[1_{N_1} \operatorname{tr}\left(\mathbf{U}' R_K (R_K' R_K)^{-1} R_K' \mathbf{U}\right) \middle| X\right]/N_1 \\
&= \mathbb{E}\left[1_{N_1} \operatorname{tr}\left(R_K \left(R_K' R_K\right)^{-1} R_K' \mathbf{U}\mathbf{U}'\right) \middle| X\right]/N_1 \\
&= \operatorname{tr}\left(R_K \left(R_K' R_K\right)^{-1} R_K' \mathbb{E}\left[1_{N_1} \mathbf{U}\mathbf{U}' \middle| X\right]\right)/N_1 \\
&\leq \bar{\sigma}^2 \cdot \operatorname{tr}\left(R_K (R_K' R_K)^{-1} R_K'\right)/N_1 \\
&= \bar{\sigma}^2 \cdot K/N_1
\end{aligned}
$$

Then by the Markov inequality $1_{N_1}|\hat{\Omega}_{K,N_1}^{-1/2} R_K' \mathbf{U}/N_1| = O_p(K^{1/2}/N^{-1/2})$.

Next, consider part (ii). Using lemma A.1(v),

$$
1_{N_1} \cdot \left|\hat{\Omega}_{K,N_1}^{-1} R_K' \mathbf{U}/N_1\right| \leq 1_{N_1} \cdot \lambda_{\max}(\hat{\Omega}_{K,N_1}^{-1/2}) \cdot \left|\hat{\Omega}_{K,N_1}^{-1/2} R_K' \mathbf{U}/N_1\right|.
$$

Since $1_{N_1} \cdot \lambda_{\max}(\hat{\Omega}_{K,N_1}^{-1/2}) = O_p(1)$, the conclusion follows. □

The conditional expectation $\mu(x)$ is estimated by the series estimator $\hat{\mu}_K(x) = R_K'(x)\hat{\gamma}_K$ with $\hat{\gamma}_K$ the least squares estimator. Formally $R_K' R_K$ may be singular, although lemma A.3 shows that this happens with probability going to zero. To deal with this case we define $\hat{\gamma}_K$ as

$$
\hat{\gamma}_K = \begin{cases} 0 & \text{if } \lambda_{\min}(\hat{\Omega}_{K,N_1}) \leq 1/2, \\ \left(\sum_{i=1}^{N_1} R_K(X_i) R_K'(X_i)\right)^{-1} \sum_{i=1}^{N_1} R_K(X_i) Y_i & \text{otherwise,} \end{cases} \tag{A.1}
$$

By lemma A.3, it follows that a sufficient condition for $1_{N_1} \xrightarrow{p} 1$ is $\zeta(K)K^{1/2}N^{-1/2} \to 0$, i.e. K increases with N at a rate that ensures that $\hat{\Omega}_{K,N_1} \xrightarrow{p} I_K$. This implies that for any sequence a_N we have $a_N 1_{N_1} \xrightarrow{p} 0$ so that if the rate condition is met, the indicator does not affect the rate of convergence of the OLS estimator.

Define γ_K^* to be the pseudo true value defined as

$$
\gamma_K^* = \arg\min_{\gamma} \mathbb{E}\left[\left(\mu(X) - R_K(X)'\gamma\right)^2 \middle| W = 1\right], \tag{A.2}
$$

with the corresponding pseudo true value of the regression function denoted by $\mu_K^*(x) = R_K(x)'\gamma_K^*$.

First we state some of the properties of the estimator for the regression function. In order to do so it is useful to first give some approximation properties for $\mu(x)$.

Assumption A.3 *$\mu(x)$ is s_μ times continuously differentiable on $\mathbb{X}$.*

This assumption together with assumption A.1 implies that μ is bounded on $\mathbb{X}$.

Lemma A.5 (Lorentz, 1986) *Suppose assumptions A.1 and A.3 hold. Then there is a sequence γ_K^0 such that*

$$
\sup_{x \in \mathbb{X}} |\mu(x) - R_K(x)'\gamma_K^0| = O\left(K^{-s_\mu/d}\right).
$$

For the sequence γ_K^0 in lemma A.5, define the corresponding sequence of regression functions, $\mu_K^0(x) = R_K(x)'\gamma_K^0$.

Lemma A.6 (Convergence Rate for Regression Function Estimators)
Suppose assumptions A.1-A.3 hold. Then (i)*:*

$$\left|\gamma_K^* - \gamma_K^0\right| = O\left(K^{1/2-s_\mu/d}\right), \tag{A.3}$$

(ii)*:*

$$\sup_{x\in\mathbb{X}} |\mu_K^*(x) - \mu_K^0(x)| = O(\zeta(K)K^{1/2-s_\mu/d}). \tag{A.4}$$

(iii)*:*

$$|\hat{\gamma}_K - \gamma_K^*| = O_p(\zeta(K)KN^{-1/2}), \tag{A.5}$$

(iv)*:*

$$\left|\hat{\gamma}_K - \gamma_K^0\right| = O_p\left(\zeta(K)K^{1/2}N^{-1/2} + K^{-s_\mu/d}\right), \tag{A.6}$$

(v)*:*

$$\sup_{x\in\mathbb{X}} |\hat{\mu}_K(x) - \mu_K^*(x)| = O_p(\zeta(K)^2KN^{-1/2}), \tag{A.7}$$

and (vi)*:*

$$\sup_{x\in\mathbb{X}} |\hat{\mu}_K(x) - \mu(x)| = O_p(\zeta(K)^2K^{1/2}N^{-1/2} + \zeta(K)K^{-s_\mu/d})), \tag{A.8}$$

Proof: First, consider (i):

$$\begin{aligned}\gamma_K^* &= \left(\mathbb{E}[R_K(X)R_K'(X)]\right)^{-1}\mathbb{E}[R_K(X)Y] = \left(\mathbb{E}[R_K(X)R_K'(X)]\right)^{-1}\mathbb{E}[R_K(X)\mu(X)] \\ &= \mathbb{E}[R_K(X)\mu(X)],\end{aligned}$$

where we use $\mathbb{E}[R_K(X)R_K(X)'] = I_K$. Also, $\gamma_K^0 = \mathbb{E}[R_K(X)R_K'(X)\gamma_K^0]$, so that

$$\begin{aligned}|\gamma_K^* - \gamma_K^0| &= |\mathbb{E}[R_K(X)\mu(X)] - \mathbb{E}[R_K(X)R_K(X)'\gamma_K^0]| \\ &= |\mathbb{E}[R_K(X)(\mu(X) - R_K(X)'\gamma_K^0)]| \leq \mathbb{E}[|R_K(X)||\mu(X) - R_K(X)'\gamma_K^0|] \\ &\leq \sqrt{\mathbb{E}\left[(\mu(X) - R_K(X)'\gamma_K^0)^2\right]\mathbb{E}\left[R_K(X)'R_K(X)\right]} \\ &\leq \sqrt{\sup_{x\in\mathbb{X}}(\mu(x) - R_K(x)'\gamma_K^0)^2 \cdot \mathbb{E}\left[R_K(X)'R_K(X)\right]} = O\left(K^{1/2}K^{-s_\mu/d}\right).\end{aligned}$$

because $\mathbb{E}[R_X(X)'R_K(X)] = \mathrm{tr}(\mathbb{E}[R_X(X)R_K(X)']) = K$. Next, consider (ii):

$$\begin{aligned}\sup_{x\in\mathbb{X}} |\mu_K^*(x) - \mu_K^0(x)| &= \sup_{x\in\mathbb{X}} |R_K(x)'(\gamma_K^* - \gamma_K^0)| \\ &\leq \sup_{x\in\mathbb{X}} |R_K(x)| \cdot |\gamma_K^* - \gamma_K^0| = O\left(\zeta(K)K^{1/2-s_\mu/d}\right),\end{aligned}$$

using the result in lemma A.6(i).

Next, consider (iii): The bound that is derived is proportional to $N^{-1/2}$, so that it converges to 0 for fixed K. Using the same method of proof as in (iv) below, a 'better' bound can be obtained that however only converges to 0 if $K \to \infty$ as well. We have

$$|\hat{\gamma}_K - \gamma_K^*| = 1_{N_1} \cdot |\hat{\gamma}_K - \gamma_K^*| + (1 - 1_{N_1}) \cdot |\gamma_K^*|,$$

The second term $(1-1_{N_1})\cdot|\gamma_K^*|$ is $o_p(\zeta(K)K^{1/2}N^{-1/2})$, in the sense that it is nonzero only if $\lambda_{\min}(\hat{\Omega}_{K,N_1}) \leq 1/2$ and the probability of this event converges to 0 if the rate condition is met[5]. The first term is

$$\begin{aligned}
1_{N_1}\cdot|\hat{\gamma}_K - \gamma_K^*| &= 1_{N_1}\cdot|\hat{\Omega}_{K,N_1}^{-1}(R_K'\mathbf{Y}/N_1) - \hat{\Omega}_{K,N_1}^{-1}(R_K'R_K\gamma_K^*/N_1)| \\
&= 1_{N_1}\cdot|\hat{\Omega}_{K,N_1}^{-1}(R_K'\mathbf{U}/N_1) + \hat{\Omega}_{K,N_1}^{-1}(R_K'(\mu(\mathbf{X}) - R_K\gamma_K^*)/N_1)| \\
&\leq 1_{N_1}\cdot|\hat{\Omega}_{K,N_1}^{-1}(R_K'\mathbf{U}/N_1)| + 1_{N_1}\cdot|\hat{\Omega}_{K,N_1}^{-1}(R_K'(\mu(\mathbf{X}) - R_K\gamma_K^*)/N_1)|.
\end{aligned}$$

By lemma A.4(ii) the first term is $O_p(K^{1/2}N^{-1/2})$. For the second term, we have, using lemma A.1(v), and using the fact that if 1_{N_1} is equal to 1, then $\lambda_{\max}(\hat{\Omega}_{K,N_1}^{-1}) \leq 2$,

$$1_{N_1}\cdot|\hat{\Omega}_{K,N_1}^{-1}(R_K'(\mu(\mathbf{X}) - R_K\gamma_K^*)/N_1)| \leq 1_{N_1}\cdot 2|(R_K'(\mu(\mathbf{X}) - R_K\gamma_K^*)/N_1)|$$

Now,

$$R_K'(\mu(\mathbf{X}) - R_K\gamma_K^*)/N_1 = \frac{1}{N_1}\sum_{i=1}^{N_1}(R_K(X_i)\mu(X_i) - R_K(X_i)R_K(X_i)'\gamma_K^*)$$

which is an average of mean 0 random variables. By the Markov inequality the rate of convergence is the square root of

$$\frac{1}{N_1}\mathbb{E}\left[(R_K(X)\mu(X) - R_K(X)R_K(X)'\gamma_K^*)'(R_K(X)\mu(X) - R_K(X)R_K(X)'\gamma_K^*)\right]$$

$$= \frac{1}{N_1}\mathbb{E}\left[\mu(X)^2R_K(X)'R_K(X) + \gamma_K^{*'}R_K(X)R_K(X)'R_K(X)R_K(X)'\gamma_K^* - 2\mu(X)'R_K(X)'R_K(X)R_K(X)'\gamma_K^*\right]$$

The first term of the final expression is bounded by

$$\frac{\sup_{x\in\mathbb{X}}\mu(x)^2}{N_1}\mathbb{E}[R_K(X)'R_K(X)] \leq C_1\frac{K}{N}$$

the second by using lemma A.1(ii) and $\zeta(K) = \sup_{x\in\mathbb{X}}|R_K(x)|$ is bounded by

$$\frac{|\gamma_K^*|^2}{N_1}\mathbb{E}[|R_K(X)R_K(X)'R_K(X)R_K(X)'|] = \frac{|\gamma_K^*|^2}{N_1}\mathbb{E}[|R_K(X)|^2|R_K(X)'R_K(X)|] \leq C_2\frac{\zeta(K)^2K^2}{N}$$

because $|\gamma_K^*| = |\mathbb{E}[R_K(X)\mu(X)]| \leq \sup_{x\in\mathbb{X}}|\mu(x)|\sqrt{\mathbb{E}[R_K(X)'R_K(X)]} = O(\sqrt{K})$, and the third term is bounded by

$$2\frac{\sup_{x\in\mathbb{X}}|\mu(x)|\cdot\sup_{x\in\mathbb{X}}|R_K(x)||\gamma_K^*|}{N_1}\mathbb{E}[R_K(X)'R_K(X)] \leq C_3\frac{\zeta(K)K^{3/2}}{N}$$

Hence

$$|\hat{\gamma}_K - \gamma_K^*| = O_p(\zeta(K)KN^{-1/2})$$

Next, consider (iv): In this case the bound requires that both $K, N \to \infty$. We have

$$|\hat{\gamma}_K - \gamma_K^0| = 1_{N_1}\cdot|\hat{\gamma}_K - \gamma_K^0| + (1-1_{N_1})|\gamma_K^0|$$

[5]This is an abuse of notation because it only expresses that the term is $o_p(1)$ if the rate condition holds, but the usual rules for stochastic order terms do not apply.

The second term is $o_p(\zeta(K)K^{1/2}N^{-1/2})$ as was the analogous term in the proof of (iii). Also

$$1_{N_1} \cdot |\hat{\gamma}_K - \gamma_K^0| = 1_{N_1} \cdot |\hat{\Omega}_{K,N_1}^{-1}(R_K'\mathbf{Y}/N_1) - \hat{\Omega}_{K,N_1}^{-1}(R_K'R_K\gamma_K^0/N_1)|$$

$$\leq 1_{N_1} \cdot |\hat{\Omega}_{K,N_1}^{-1}(R_K'\mathbf{U}/N_1)| + 1_{N_1} \cdot |\hat{\Omega}_{K,N_1}^{-1}(R_K'(\mu(\mathbf{X}) - R_K'\gamma_K^0)/N_1)|.$$

By lemma A.4(ii) the first term is $O_p(K^{1/2}N^{-1/2})$. For the second term, we have, using lemma A.1(v), and using the fact that if 1_{N_1} is equal to 1, then $\lambda_{\max}(\hat{\Omega}_{K,N_1}^{-1/2}) \leq \sqrt{2}$,

$$1_{N_1} \cdot |\hat{\Omega}_{K,N_1}^{-1}(R_K'(\mu(\mathbf{X}) - R_K\gamma_K^0)/N_1)| \leq 1_{N_1} \cdot \lambda_{\max}(\hat{\Omega}_{K,N_1}^{-1/2}) \cdot |\hat{\Omega}_{K,N_1}^{-1/2}(R_K'(\mu(\mathbf{X}) - R_K\gamma_K^0)/N_1)|$$

$$\leq 1_{N_1} \cdot \sqrt{2} \cdot \left(\frac{1}{N_1^2}(\mu(\mathbf{X}) - R_K\gamma_K^0)'R_K'\hat{\Omega}_{K,N_1}^{-1/2}\hat{\Omega}_{K,N_1}^{-1/2}R_K'(\mu(\mathbf{X}) - R_K\gamma_K^0)\right)^{1/2}$$

$$= 1_{N_1} \cdot \sqrt{2} \cdot \left(\frac{1}{N_1}(\mu(\mathbf{X}) - R_K\gamma_K^0)'R_K(R_K'R_K)^{-1}R_K'(\mu(\mathbf{X}) - R_K\gamma_K^0)\right)^{1/2}$$

$$\leq \sqrt{2} \cdot \left(\frac{1}{N_1}(\mu(\mathbf{X}) - R_K\gamma_K^0)'(\mu(\mathbf{X}) - R_K\gamma_K^0)\right)^{1/2} \tag{A.9}$$

where we use the fact that because $R_K(R_K'R_K)^{-1}R_K'$ is a projection matrix, it follows that $I_{N_1} - R_K(R_K'R_K)^{-1}R_K'$ is positive semi-definite. Since

$$\mathbb{E}\left[\frac{1}{N_1}(\mu(\mathbf{X}) - R_K\gamma_K^0)'(\mu(\mathbf{X}) - R_K\gamma_K^0)\right]$$

$$\leq \sup_{x\in\mathbb{X}} |\mu(x) - R_K(x)'\gamma_K^0|^2 \leq CK^{-2s_\mu/d},$$

it follows by the Markov inequality that (A.9) is $O_p(K^{-s_\mu/d})$. Hence $\|\hat{\gamma}_K - \gamma_K^0\| = O_p(K^{-s_\mu/d}) + O_p(K^{1/2}N^{-1/2}) + O_p(\zeta(K)K^{1/2}N^{-1/2}) = O_p(K^{-s_\mu/d} + \zeta(K)K^{1/2}N^{-1/2})$. Note that the bound on the variance is obtained from the rate of convergence of $\hat{\Omega}_{K,N_1}$.

Next, consider (v):

$$\sup_{x\in\mathbb{X}} |\hat{\mu}_K(x) - \mu_K^*(x)| = \sup_{x\in\mathbb{X}} |R_K(x)'(\hat{\gamma}_K - \gamma_K^*)|$$

$$\leq \sup_{x\in\mathbb{X}} |R_K(x)| \cdot |\hat{\gamma}_K - \gamma_K^*| = O_p(\zeta(K)^2KN^{-1/2}).$$

Finally, consider (vi).

$$\sup_{x\in\mathbb{X}} |\hat{\mu}_K(x) - \mu(x)| \leq \sup_{x\in\mathbb{X}} |\hat{\mu}_K(x) - \mu_K^0(x)| + \sup_{x\in\mathbb{X}} |\mu_K^0(x) - \mu(x)|$$

$$= O_p(\zeta(K)^2K^{1/2}N^{-1/2} + \zeta(K)K^{-s_\mu/d} + K^{-s_\mu/d}) = O_p(\zeta(K)^2K^{1/2}N^{-1/2} + \zeta(K)K^{-s_\mu/d}).$$

□

Here we briefly summarize the relevant results from HIR for the nonparametric estimator for the propensity score. Let $L(z) = \exp(z)/(1+\exp(z))$ be the logistic cdf and $L'(z) = L(z) \cdot (1 - L(z))$. The series logit estimator of the population propensity score $e(x)$ is $\hat{e}_L(x) = L(R_K(x)'\hat{\pi}_L)$, where

$$\hat{\pi}_L = \arg\max_{\pi} L_N(\pi), \tag{A.10}$$

for

$$L_N(\pi) = \sum_{i=1}^{N} \left(W_i \cdot \ln L(R_L(X_i)'\pi) + (1 - W_i) \cdot \ln(1 - L(R_L(X_i)'\pi))\right). \tag{A.11}$$

For $N \to \infty$ and L fixed we have $\hat{\pi}_L \xrightarrow{p} \pi_L^*$, with π_L^* the pseudo true value:

$$\pi_L^* = \arg\max_{\pi} \mathbb{E}\left[e(X) \cdot \ln L(R_L(X)'\pi) + (1 - e(X)) \cdot \ln(1 - L(R_L(X)'\pi))\right]. \tag{A.12}$$

We also define the pseudo true propensity score: $e_L^*(x) = L(R_L(x)'\pi_L^*)$.

Assumption A.4 *$e(x)$ is s_e times continuously differentiable on $\mathbb{X}$.*

Lemma A.7 *Suppose assumptions A.1 and A.4 hold. Then there is a sequence π_L^0 such that*

$$\sup_{x \in \mathbb{X}} |e(x) - L(R_L(x)'\pi_L^0)| = O\left(L^{-s_e/d}\right).$$

Proof: See HIR.

We define $e_L^0(x) = L(R_L(x)'\pi_L^0)$.

Assumption A.5 $\inf_x e(x) > 0$ *and* $\sup_{x \in \mathbb{X}} e(x) < 1$.

Lemma A.8 (Convergence Rate for Propensity Score Estimators)
Suppose assumptions A.1, A.4 and A.5 hold and $s_e/d \geq 4$.

Then (i):

$$\left|\left|\pi_L^* - \pi_L^0\right|\right| = O\left(L^{-s_e/(2d)}\right), \tag{A.13}$$

(ii):

$$\sup_{x \in \mathbb{X}} \left|e_L^*(x) - e_L^0(x)\right| = O\left(\zeta(L) L^{-s_e/(2d)}\right). \tag{A.14}$$

(iii):

$$||\hat{\pi}_L - \pi_L^*|| = O_p(\zeta(L)^2 L^{1/2} N^{-1/2}), \tag{A.15}$$

(iv),

$$|\hat{\pi}_L - \pi_L^0| = O_p(\zeta(L) L^{1/2} N^{-1/2} + L^{-s_e/(2d)}), \tag{A.16}$$

(v):

$$\sup_{x \in \mathbb{X}} |\hat{e}_L(x) - e_L^*(x)| = O_p\left(\zeta(L)^3 L^{1/2} N^{-1/2}\right). \tag{A.17}$$

(vi),

$$\sup_{x \in \mathbb{X}} |\hat{e}_L(x) - e_L^0(x)| = O_p(\zeta(L)^2 L^{1/2} N^{-1/2} + \zeta(L) L^{-s_e/(2d)}), \tag{A.18}$$

and (vii),

$$\sup_{x \in \mathbb{X}} |\hat{e}_L(x) - e(x)| = O_p(\zeta(L)^2 L^{1/2} N^{-1/2} + \zeta(L) L^{-s_e/(2d)})). \tag{A.19}$$

Proof: See HIR.
Note that the bound in (ii) of lemma A.8 is not faster than that in (ii) of lemma A.7, if μ and e are equally smooth and $s_\mu/d = s_e/d \geq 1$.

The second estimator is a modified imputation estimator that only averages over the observations with $W_i = 1$:

$$\hat{\tau}_{mod} = \frac{1}{N}\sum_{i=1}^{N} \frac{W_i \cdot \hat{\mu}_K(X_i)}{\hat{e}_L(X_i)}. \tag{A.20}$$

The third estimator is the weighting estimator proposed by HIR:

$$\hat{\tau}_{hir,1} = \frac{1}{N}\sum_{i=1}^{N} \frac{W_i \cdot Y_i}{\hat{e}_L(X_i)}. \tag{A.21}$$

HIR show that $\hat{\tau}_{hir,1}$ is consistent, asymptotically normal and efficient.

The fourth estimator is a modified version of the Hirano-Imbens-Ridder estimator where the weights are normalized to add up to unity:

$$\hat{\tau}_{hir,2} = \sum_{i=1}^{N} \frac{W_i \cdot Y_i}{\hat{e}_L(X_i)} \Big/ \sum_{i=1}^{N} \frac{W_i}{\hat{e}_L(X_i)}. \tag{A.22}$$

The properties of $\hat{\tau}_{imp}$ follow from theorem 3.1 that establishes that all estimators are asymptotically equivalent.

Proof of theorem 3.1:

First we prove (i). In this proof (and the proofs of the other parts) we encounter four types of terms that have distinct asymptotic bounds:

a. Terms with an upper bound that depends solely on the smoothness of the conditional mean function and/or the propensity score. These terms involve the cross-product of an estimate and the bias of either the conditional mean function or the propensity score and are bounded by lemmas A.5 and A.7

b. Terms with an upper bound as in lemma A.6, (v), (vi), or lemma A.8, (vi), (vii). These terms involve the cross-product of estimates of either the conditional mean function and the propensity score.

c. Terms that are degenerate U-statistics.

d. Terms that are projection remainders.

We have

$$\sqrt{N}|\hat{\tau}_{imp} - \hat{\tau}_{mod}| = \left| \frac{1}{\sqrt{N}}\sum_{i=1}^{N} \hat{\mu}_K(X_i) - \frac{1}{\sqrt{N}}\sum_{i=1}^{N} \frac{W_i \cdot \hat{\mu}_K(X_i)}{\hat{e}_L(X_i)} \right| = \left| \frac{1}{\sqrt{N}}\sum_{i=1}^{N} \left(\frac{\hat{\mu}_K(X_i) \cdot \hat{e}_L(X_i)}{\hat{e}_L(X_i)} - \frac{W_i \cdot \hat{\mu}_K(X_i)}{\hat{e}_L(X_i)} \right) \right| \tag{A.23}$$

$$= \left| \frac{1}{\sqrt{N}}\sum_{i=1}^{N} \left\{ \hat{\mu}_K(X_i) \left(\frac{1}{\hat{e}_L(X_i)} - \frac{1}{e_L^0(X_i)} \right) + \frac{1}{e_L^0(X_i)} \left(\hat{\mu}_K(X_i) - \mu_K^0(X_i) \right) \right.$$

$$+\mu_K^0(X_i)\left(\frac{1}{e_L^0(X_i)}-\frac{1}{e(X_i)}\right)+\frac{1}{e(X_i)}\left(\mu_K^0(X_i)-\mu(X_i)\right)\Big\}\left(\hat{e}_L(X_i)-W_i\right)+\frac{\mu(X_i)}{e(X_i)}(\hat{e}_L(X_i)-W_i)\Bigg|$$

$$\leq\left|\frac{1}{\sqrt{N}}\sum_{i=1}^{N}\hat{\mu}_K(X_i)\left(\frac{1}{\hat{e}_L(X_i)}-\frac{1}{e_L^0(X_i)}\right)\left(\hat{e}_L(X_i)-W_i\right)\right| \tag{A.24}$$

$$+\left|\frac{1}{\sqrt{N}}\sum_{i=1}^{N}\frac{1}{e_L^0(X_i)}\left(\hat{\mu}_K(X_i)-\mu_K^0(X_i)\right)\left(\hat{e}_L(X_i)-W_i\right)\right| \tag{A.25}$$

$$+\left|\frac{1}{\sqrt{N}}\sum_{i=1}^{N}\mu_K^0(X_i)\left(\frac{1}{e_L^0(X_i)}-\frac{1}{e(X_i)}\right)\left(\hat{e}_L(X_i)-W_i\right)\right| \tag{A.26}$$

$$+\left|\frac{1}{\sqrt{N}}\sum_{i=1}^{N}\frac{1}{e(X_i)}\left(\mu_K^0(X_i)-\mu(X_i)\right)\cdot\left(\hat{e}_L(X_i)-W_i\right)\right| \tag{A.27}$$

$$+\left|\frac{1}{\sqrt{N}}\sum_{i=1}^{N}\frac{\mu(X_i)}{e(X_i)}(\hat{e}_L(X_i)-W_i)\right|. \tag{A.28}$$

We will deal with (A.24)-(A.28) separately. First consider (A.24).

$$\left|\frac{1}{\sqrt{N}}\sum_{i=1}^{N}\hat{\mu}_K(X_i)\cdot\left(\frac{1}{\hat{e}_L(X_i)}-\frac{1}{e_L^0(X_i)}\right)\cdot\left(\hat{e}_L(X_i)-W_i\right)\right|$$

$$\leq\left|\frac{1}{\sqrt{N}}\sum_{i=1}^{N}\left(\hat{\mu}_K(X_i)-\mu_K^0(X_i)\right)\cdot\left(\frac{1}{\hat{e}_L(X_i)}-\frac{1}{e_L^0(X_i)}\right)\cdot\left(\hat{e}_L(X_i)-W_i\right)\right| \tag{A.29}$$

$$+\left|\frac{1}{\sqrt{N}}\sum_{i=1}^{N}\left(\mu_K^0(X_i)-\mu(X_i)\right)\cdot\left(\frac{1}{\hat{e}_L(X_i)}-\frac{1}{e_L^0(X_i)}\right)\cdot\left(\hat{e}_L(X_i)-W_i\right)\right| \tag{A.30}$$

$$+\left|\frac{1}{\sqrt{N}}\sum_{i=1}^{N}\mu(X_i)\cdot\left(\frac{1}{\hat{e}_L(X_i)}-\frac{1}{e_L^0(X_i)}\right)\cdot\left(\hat{e}_L(X_i)-W_i\right)\right| \tag{A.31}$$

First consider (A.29). Since $\inf_{x\in\mathbb{X}}e(x)>c$ for some $c>0$, it follows from $e_L^0(x)\geq e(x)-\sup_{x\in\mathbb{X}}|e(x)-e_L^0(x)|$ and lemma A.7 that if N (and hence L) is sufficiently large $\inf_{x\in\mathbb{X}}e_L^0(x)>c/2$. Also analogously by lemma A.8, (vii) for large enough N (and hence[6] L), with arbitrarily high probability , $\inf_x\hat{e}_L(x)>c/2$. Thus by lemma A.8, (vi)

$$\sup_{x\in\mathbb{X}}\left|\frac{1}{\hat{e}_L(x)}-\frac{1}{e_L^0(x)}\right|=\sup_{x\in\mathbb{X}}\frac{1}{e_L^0(x)\hat{e}_L(x)}\left|\hat{e}_L(x)-e_L^0(x)\right|=O_p(\zeta(L)^2L^{-s_e/(2d)}+\zeta(L)L^{1/2}N^{-1/2}).$$

Thus by lemma A.5 and A.6 (vi)

$$\left|\frac{1}{\sqrt{N}}\sum_{i=1}^{N}\left(\hat{\mu}_K(X_i)-\mu_K^0(X_i)\right)\cdot\left(\frac{1}{\hat{e}_L(X_i)}-\frac{1}{e_L^0(X_i)}\right)\cdot\left(\hat{e}_L(X_i)-W_i\right)\right|$$

$$\leq N^{1/2}\cdot\sup_{x\in\mathbb{X}}\left|\hat{\mu}_K(x)-\mu_K^0(x)\right|\cdot\sup_{x\in\mathbb{X}}\left|\frac{1}{\hat{e}_L(x)}-\frac{1}{e_L^0(x)}\right|\cdot 2$$

[6]We require that $\zeta(L)^2L^{1/2}N^{-1/2}+\zeta(L)L^{-s_e/(2d)}\to 0$.

$$= O_p\left(N^{1/2} \cdot (\zeta(K)K^{-s_\mu/d} + \zeta(K)^2K^{1/2}N^{-1/2}) \cdot (\zeta(L)L^{-s_e/(2d)} + \zeta(L)^2L^{1/2}N^{-1/2})\right).$$

Note that this is a bound of type b.

Next, consider (A.30). Using the bound derived above and lemma A.5

$$\left|\frac{1}{\sqrt{N}}\sum_{i=1}^{N}\left(\mu_K^0(X_i) - \mu(X_i)\right) \cdot \left(\frac{1}{\hat{e}_L(X_i)} - \frac{1}{e_L^0(X_i)}\right) \cdot (\hat{e}_L(X_i) - W_i)\right|$$

$$\leq N^{1/2} \cdot \sup_{x\in\mathbb{X}}|\mu_K^0(x) - \mu(x))| \cdot \sup_{x\in\mathbb{X}}\left|\frac{1}{\hat{e}_L(X_i)} - \frac{1}{e_L^0(X_i)}\right| \cdot 2$$

$$= O_p\left(N^{1/2}K^{-s_\mu/d}(\zeta(L)L^{-s_e/(2d)} + \zeta(L)^2L^{1/2}N^{-1/2})\right).$$

This is a bound of type a.

Finally, consider (A.31).

$$\left|\frac{1}{\sqrt{N}}\sum_{i=1}^{N}\mu(X_i) \cdot \left(\frac{1}{\hat{e}_L(X_i)} - \frac{1}{e_L^0(X_i)}\right) \cdot (\hat{e}_L(X_i) - W_i)\right|$$

$$\leq \left|\frac{1}{\sqrt{N}}\sum_{i=1}^{N}\mu(X_i) \cdot \frac{1}{\hat{e}_L(X_i)e_L^0(X_i)}\left(\hat{e}_L(X_i) - e_L^0(X_i)\right) \cdot (\hat{e}_L(X_i) - e(X_i))\right| \tag{A.32}$$

$$+ \left|\frac{1}{\sqrt{N}}\sum_{i=1}^{N}\mu(X_i) \cdot \frac{1}{\hat{e}_L(X_i)e_L^0(X_i)}\left(\hat{e}_L(X_i) - e_L^0(X_i)\right) \cdot (e(X_i) - W_i)\right| \tag{A.33}$$

For (A.32) we have by lemma A.8 (vi) and (vii)

$$\left|\frac{1}{\sqrt{N}}\sum_{i=1}^{N}\mu(X_i) \cdot \frac{1}{\hat{e}_L(X_i)e_L^0(X_i)}\left(\hat{e}_L(X_i) - e_L^0(X_i)\right) \cdot (\hat{e}_L(X_i) - e(X_i))\right|$$

$$\leq \sqrt{N}\sup_{x\in\mathbb{X}}\frac{\mu(x)}{\hat{e}_L(x)e_L^0(x)}\sup_{x\in\mathbb{X}}|\hat{e}_L(x) - e_L^0(x)|\sup_{x\in\mathbb{X}}|\hat{e}_L(x) - e(x)| = O_p\left(N^{1/2}(\zeta(L)^2L^{1/2}N^{-1/2} + \zeta(L)L^{-s_e/(2d)})^2\right)$$

a type b bound. For (A.33)

$$\left|\frac{1}{\sqrt{N}}\sum_{i=1}^{N}\mu(X_i) \cdot \frac{1}{\hat{e}_L(X_i)e_L^0(X_i)}\left(\hat{e}_L(X_i) - e_L^0(X_i)\right) \cdot (e(X_i) - W_i)\right|$$

$$\leq \left|\frac{1}{\sqrt{N}}\sum_{i=1}^{N}\mu(X_i) \cdot \frac{1}{\hat{e}_L(X_i)e_L^0(X_i)}\left(\hat{e}_L(X_i) - e_L^*(X_i)\right) \cdot (e(X_i) - W_i)\right| \tag{A.34}$$

$$+ \left|\frac{1}{\sqrt{N}}\sum_{i=1}^{N}\mu(X_i) \cdot \frac{1}{\hat{e}_L(X_i)e_L^0(X_i)}\left(e_L^*(X_i) - e_L^0(X_i)\right) \cdot (e(X_i) - W_i)\right| \tag{A.35}$$

with using lemma A.8 (ii) for (A.35) the type a bound

$$\left|\frac{1}{\sqrt{N}}\sum_{i=1}^{N}\mu(X_i) \cdot \frac{1}{\hat{e}_L(X_i)e_L^0(X_i)}\left(e_L^*(X_i) - e_L^0(X_i)\right) \cdot (e(X_i) - W_i)\right| \leq \sqrt{N}\sup_{x\in\mathbb{X}}\frac{\mu(x)}{\hat{e}_L(x)e_L^0(x)}\sup_{x\in\mathbb{X}}|e_L^*(x) - e_L^0(x)| \cdot 2$$

$$= O_p\left(N^{1/2}\zeta(L)L^{-s_e/(2d)}\right)$$

For (A.34)

$$\left|\frac{1}{\sqrt{N}}\sum_{i=1}^{N}\mu(X_i)\cdot\frac{1}{\hat{e}_L(X_i)e_L^0(X_i)}\left(\hat{e}_L(X_i)-e_L^*(X_i)\right)\cdot\left(e(X_i)-W_i\right)\right|$$

$$\leq\left|\frac{1}{\sqrt{N}}\sum_{i=1}^{N}\mu(X_i)\cdot\frac{1}{e_L^0(X_i)^2}\left(\hat{e}_L(X_i)-e_L^*(X_i)\right)\cdot\left(e(X_i)-W_i\right)\right| \tag{A.36}$$

$$+\left|\frac{1}{\sqrt{N}}\sum_{i=1}^{N}\mu(X_i)\cdot\frac{1}{\hat{e}_L(X_i)e_L^0(X_i)^2}\left(\hat{e}_L(X_i)-e_L^0(X_i)\right)\cdot\left(\hat{e}_L(X_i)-e_L^*(X_i)\right)\cdot\left(e(X_i)-W_i\right)\right| \tag{A.37}$$

with by lemma A.8 (v) and (vi) for (A.37) the type b bound

$$\left|\frac{1}{\sqrt{N}}\sum_{i=1}^{N}\mu(X_i)\cdot\frac{1}{\hat{e}_L(X_i)e_L^0(X_i)^2}\left(\hat{e}_L(X_i)-e_L^0(X_i)\right)\cdot\left(\hat{e}_L(X_i)-e_L^*(X_i)\right)\cdot\left(e(X_i)-W_i\right)\right|$$

$$\leq\sqrt{N}\sup_{x\in\mathbb{X}}\frac{\mu(x)}{\hat{e}_L(x)e_L^0(x)^2}\sup_{x\in\mathbb{X}}|\hat{e}_L(x)-e_L^0(x)|\sup_{x\in\mathbb{X}}|\hat{e}_L(x)-e_L^*(x)|\cdot 2$$

$$=O_p\left(N^{1/2}\zeta(L)^3L^{1/2}N^{-1/2}(\zeta(L)^2L^{1/2}N^{-1/2}+\zeta(L)L^{-s_e/(2d)})\right)$$

To derive a bound on (A.36) we use the fact that $\hat{e}_L(X_i)-e_L(X_i)$ is a residual, because if we use lemma A.8 (iii) we obtain a bound that does not converge to 0. By a first-order Taylor series expansions of the residual and the likelihood equation we have

$$\hat{e}_L(X_i)-e_L^*(X_i)=L'(R_L(X_i)'\tilde{\pi}_L)R_L(X_i)'(\hat{\pi}_L-\pi_L^*)$$

$$=L'(R_L(X_i)'\tilde{\pi}_L)R_L(X_i)'\left(\frac{1}{N}\sum_{j=1}^{N}L'(R_L(X_j)'\overline{\pi}_L)R_L(X_j)R_L(X_j)'\right)^{-1}\left(\frac{1}{N}\sum_{j=1}^{N}(W_j-e_L^*(X_j))R_L(X_j)\right) \tag{A.38}$$

with $\tilde{\pi}_L$ and $\overline{\pi}_L$ intermediate between $\hat{\pi}_L$ and π_L^*. Define

$$\hat{\Sigma}_{L,N}\equiv\frac{1}{N}\sum_{j=1}^{N}L'(R_L(X_j)'\overline{\pi}_L)R_L(X_j)R_L(X_j)'$$

and

$$\Sigma_L\equiv\mathbb{E}\left[e_L^*(X)(1-e_L^*(X))R_L(X)R_L(X)'\right]$$

We need that $\hat{\Sigma}_{L,N}$ is nonsingular for the series logit estimator to exist. As noted before, if $\zeta(L)^2L^{1/2}N^{-1/2}+\zeta(L)L^{-s_e/2d}\to 0$ then by lemma A.8, (vii) $\inf_{x\in\mathbb{X}}L'(R_L(x)'\overline{\pi}_L)\geq c>0$ with probability arbitrarily close to 1, if N is sufficiently large. Hence because

$$\lambda_{\min}(\hat{\Sigma}_{L,N})=\min_{a'a=1}a'\hat{\Sigma}_{L,N}a$$

and with probability arbitrarily close to 1 if N is sufficiently large, for all L vectors a, $a'\hat{\Sigma}_{L,N}a\geq ca'\hat{\Omega}_{L,N}a$, we have

$$\lambda_{\min}(\hat{\Sigma}_{L,N})\geq c\lambda_{\min}(\hat{\Omega}_{L,N})$$

Hence existence and its effect on the rate of convergence can be handled as for the nonparametric regression estimator, because existence of the regression estimator implies existence of the series logit estimator. Hence we can ignore existence if $\zeta(L)^2 L^{1/2} N^{-1/2} + \zeta(L) L^{-s_e/2d} \to 0$.

Substitution of (A.38) in (A.36) gives

$$\left| \frac{1}{\sqrt{N}} \sum_{i=1}^{N} \frac{\mu(X_i)}{e_L^0(X_i)^2} (e(X_i) - W_i) L'(R_L(X_i)'\tilde{\pi}_L) R_L(X_i)' \hat{\Sigma}_{L,N}^{-1} \left(\frac{1}{N} \sum_{j=1}^{N} (W_j - e_L^*(X_j)) R_L(X_j) \right) \right|$$

$$\leq \left| \frac{1}{\sqrt{N}} \sum_{i=1}^{N} \frac{\mu(X_i)}{e_L^0(X_i)^2} (e(X_i) - W_i) L'(R_L(X_i)'\tilde{\pi}_L) R_L(X_i)' (\hat{\Sigma}_{L,N}^{-1} - \Sigma_L^{-1}) \left(\frac{1}{N} \sum_{j=1}^{N} (W_j - e_L^*(X_j)) R_L(X_j) \right) \right| \tag{A.39}$$

$$+ \left| \frac{1}{\sqrt{N}} \sum_{i=1}^{N} \frac{\mu(X_i)}{e_L^0(X_i)^2} (e(X_i) - W_i) L'(R_L(X_i)'\tilde{\pi}_L) R_L(X_i)' \Sigma_L^{-1} \left(\frac{1}{N} \sum_{j=1}^{N} (W_j - e_L^*(X_j)) R_L(X_j) \right) \right| \tag{A.40}$$

with for (A.39)

$$\left| \frac{1}{\sqrt{N}} \sum_{i=1}^{N} \frac{\mu(X_i)}{e_L^0(X_i)^2} (e(X_i) - W_i) L'(R_L(X_i)'\tilde{\pi}_L) R_L(X_i)' (\hat{\Sigma}_L^{-1} - \Sigma_L^{-1}) \left(\frac{1}{N} \sum_{j=1}^{N} (W_j - e_L^*(X_j)) R_L(X_j) \right) \right|$$

$$\leq \left| \frac{1}{\sqrt{N}} \sum_{i=1}^{N} \frac{\mu(X_i)}{e_L^0(X_i)^2} (e(X_i) - W_i) L'(R_L(X_i)'\tilde{\pi}_L) R_L(X_i)' \right| \left| \hat{\Sigma}_L^{-1} - \Sigma_L^{-1} \right| \left| \frac{1}{N} \sum_{j=1}^{N} (W_j - e_L^*(X_j)) R_L(X_j) \right|$$

The first factor is bounded by

$$\left| \frac{1}{\sqrt{N}} \sum_{i=1}^{N} \frac{\mu(X_i)}{e_L^0(X_i)^2} (e(X_i) - W_i) L'(R_L(X_i)'\pi_L^*) R_L(X_i)' \right| +$$

$$\left| \frac{1}{\sqrt{N}} \sum_{i=1}^{N} \frac{\mu(X_i)}{e_L^0(X_i)^2} (e(X_i) - W_i) (L'(R_L(X_i)'\tilde{\pi}_L) - L'(R_L(X_i)'\pi_L^*) R_L(X_i)' \right|$$

Because

$$\mathbb{E}\left[\left| \frac{1}{\sqrt{N}} \sum_{i=1}^{N} \frac{\mu(X_i)}{e_L^0(X_i)^2} (e(X_i) - W_i) L'(R_L(X_i)'\pi_L^*) R_L(X_i)' \right| \right] \leq$$

$$\sqrt{\mathbb{E}\left[\left| \frac{1}{\sqrt{N}} \sum_{i=1}^{N} \frac{\mu(X_i)}{e_L^0(X_i)^2} (e(X_i) - W_i) L'(R_L(X_i)'\pi_L^*) R_L(X_i)' \right|^2 \right]} =$$

$$\frac{1}{\sqrt{N}} \sqrt{\mathbb{E}\left[\left(\sum_{i=1}^{N} \frac{\mu(X_i)}{e_L^0(X_i)^2} (e(X_i) - W_i) L'(R_L(X_i)'\pi_L^*) R_L(X_i)' \right) \left(\sum_{i=1}^{N} \frac{\mu(X_i)}{e_L^0(X_i)^2} (e(X_i) - W_i) L'(R_L(X_i)'\pi_L^*) R_L(X_i) \right) \right]} =$$

$$\frac{1}{\sqrt{N}} \sqrt{\sum_{i=1}^{N} \mathbb{E}\left[\frac{\mu(X_i)^2}{e_L^0(X_i)^4} (e(X_i) - W_i)^2 L'(R_L(X_i)'\pi_L^*)^2 R_L(X_i)' R_L(X_i) \right]} \leq 2 \sqrt{\sup_{x \in \mathbb{X}} \frac{\mu(x)^2}{e_L^0(x)^4} \mathbb{E}[R_L(X)' R_L(X)]} = O(\sqrt{L})$$

the first term is $O_p(\sqrt{L})$ by the Markov inequality. The second term is bounded by

$$2\sqrt{N}\sup_{x\in\mathbb{X}}|L'(R_L(x)'\tilde{\pi}_L)-L'(R_L(x)'\pi_L^*)|\left|\frac{\mu(x)}{e_L^0(x)^2}\right||R_L(x)|=O_p\left(\zeta(L)^4L^{1/2}\right)$$

which is also the bound on the first factor. For the second factor we note that

$$\left|\hat{\Sigma}_{L,N}^{-1}-\Sigma_L^{-1}\right|=\left|\hat{\Sigma}_{L,N}^{-1}\right|\left|\Sigma_L^{-1}\right|\left|\hat{\Sigma}_{L,N}-\Sigma_L\right|$$

We have

$$|\hat{\Sigma}_{L,N}^{-1}|=\sqrt{\sum_{k=1}^{L}\frac{1}{\lambda_k(\hat{\Sigma}_{L,N})^2}}\leq\sqrt{L}\frac{1}{\lambda_{min}(\hat{\Sigma}_{L,N})}=O_p(\sqrt{L})$$

if the rate that ensures identification is met. An analogous argument gives $|\Sigma_L^{-1}|=O(\sqrt{L})$. Further

$$\left|\hat{\Sigma}_{L,N}-\Sigma_L\right|\leq\left|\frac{1}{N}\sum_{j=1}^{N}(L'(R_L(X_j)'\overline{\pi}_L)-L'(R_L(X_j)'\pi_L^*))R_L(X_j)R_L(X_j)'\right|+\left|\frac{1}{N}\sum_{j=1}^{N}L'(R_L(X_j)'\pi_L^*)R_L(X_j)R_L(X_j)'-\Sigma_L\right|$$

Using lemma A.8, (v) and the Markov inequality, the first term is bounded by

$$\sup_{x\in\mathbb{X}}|L'(R_L(x)'\overline{\pi}_L)-L'(R_L(x)'\pi_L^*)|\frac{1}{N}\sum_{j=1}^{N}R_L(X_j)'R_L(X_j)=O_p\left(\zeta(L)^3L^{1/2}N^{-1/2}L\right)$$

The second term is by an argument as in the proof of lemma A.3, but for basis functions that are not orthonormal $O\left(\zeta(L)^2N^{-1/2}\right)$ so that the second factor is $O_p\left(\zeta(L)^3L^{3/2}N^{-1/2}\right)$. Because for the third factor

$$\mathbb{E}\left[\left|\frac{1}{N}\sum_{j=1}^{N}(W_j-e_L^*(X_j))R_L(X_j)\right|\right]\leq\mathbb{E}\left[|(W-e_L^*(X))R_L(X)|\right]\leq2\cdot\zeta(L)$$

we have

$$\left|\frac{1}{N}\sum_{j=1}^{N}(W_j-e_L^*(X_j))R_L(X_j)\right|=O_p(\zeta(L))$$

Combining bounds

$$\left|\frac{1}{\sqrt{N}}\sum_{i=1}^{N}\frac{\mu(X_i)}{e_L^0(X_i)^2}(e(X_i)-W_i)L'(R_L(X_i)'\tilde{\pi}_L)R_L(X_i)'(\hat{\Sigma}_{L,N}^{-1}-\Sigma_L^{-1})\left(\frac{1}{N}\sum_{j=1}^{N}(W_j-e_L^*(X_j))R_L(X_j)\right)\right|=$$

$$O_p\left(\zeta(L)^5L^3N^{-1/2}\right)$$

For (A.40)

$$\left|\frac{1}{\sqrt{N}}\sum_{i=1}^{N}\frac{\mu(X_i)}{e_L^0(X_i)^2}(e(X_i)-W_i)L'(R_L(X_i)'\tilde{\pi}_L)R_L(X_i)'\Sigma_L^{-1}\left(\frac{1}{N}\sum_{j=1}^{N}(W_j-e_L^*(X_j))R_L(X_j)\right)\right|$$

$$\leq\left|\frac{1}{\sqrt{N}}\sum_{i=1}^{N}\frac{\mu(X_i)}{e_L^0(X_i)^2}(e(X_i)-W_i)L'(R_L(X_i)'\tilde{\pi}_L)R_L(X_i)'\Sigma_L^{-1}\left(\frac{1}{N}\sum_{j=1}^{N}(W_j-e(X_j))R_L(X_j)\right)\right| \quad \text{(A.41)}$$

$$+\left|\frac{1}{\sqrt{N}}\sum_{i=1}^{N}\frac{\mu(X_i)}{e_L^0(X_i)^2}(e(X_i)-W_i)L'(R_L(X_i)'\tilde{\pi}_L)R_L(X_i)'\Sigma_L^{-1}\left(\frac{1}{N}\sum_{j=1}^{N}(e(X_j)-e_L^*(X_j))R_L(X_j)\right)\right| \tag{A.42}$$

with (A.42) bounded by

$$\left|\frac{1}{\sqrt{N}}\sum_{i=1}^{N}\frac{\mu(X_i)}{e_L^0(X_i)^2}(e(X_i)-W_i)L'(R_L(X_i)'\tilde{\pi}_L)R_L(X_i)'\right||\Sigma_L^{-1}|\left|\frac{1}{N}\sum_{j=1}^{N}(e(X_j)-e_L^*(X_j))R_L(X_j)\right|$$

The first factor is $O_p\left(\zeta(L)^4L^{1/2}\right)$ as shown above. The second factor is $O(\sqrt{L})$, and the third factor by lemmas A.5, A.8(ii) and the Markov inequality

$$\left|\frac{1}{N}\sum_{j=1}^{N}(e(X_j)-e_L^*(X_j))R_L(X_j)\right|\leq\sup_{x\in\mathbb{X}}|e(x)-e_L^*(x)|\left|\frac{1}{N}\sum_{j=1}^{N}R_L(X_j)\right|=O_p\left(L^{-s_e/(2d)}\zeta(L)\right)$$

Hence (A.42) is $O_p\left(\zeta(L)^5L^{1-s_e/(2d)}\right)\big)$.

Finally (A.41) is bounded by

$$\left|\frac{1}{\sqrt{N}}\sum_{i=1}^{N}\frac{\mu(X_i)}{e_L^0(X_i)^2}(e(X_i)-W_i)(L'(R_L(X_i)'\tilde{\pi}_L)-L'(R_L(X_i)'\pi_L^*))R_L(X_i)'\Sigma_L^{-1}\left(\frac{1}{N}\sum_{j=1}^{N}(W_j-e(X_j))R_L(X_j)\right)\right|+$$

$$\left|\frac{1}{\sqrt{N}}\sum_{i=1}^{N}\frac{\mu(X_i)}{e_L^0(X_i)^2}(e(X_i)-W_i)L'(R_L(X_i)'\pi_L^*)R_L(X_i)'\Sigma_L^{-1}\left(\frac{1}{N}\sum_{j=1}^{N}(W_j-e(X_j))R_L(X_j)\right)\right|$$

The first term is bounded by

$$\left|\frac{1}{\sqrt{N}}\sum_{i=1}^{N}\frac{\mu(X_i)}{e_L^0(X_i)^2}(e(X_i)-W_i)(L'(R_L(X_i)'\tilde{\pi}_L)-L'(R_L(X_i)'\pi_L^*))R_L(X_i)'\right||\Sigma_L^{-1}|\left|\frac{1}{N}\sum_{j=1}^{N}(W_j-e(X_j))R_L(X_j)\right|$$

with the first factor bounded by

$$2\sup_{x\in\mathbb{X}}\frac{|\mu(x)|}{e_L^0(x)^2}\sup_{x\in\mathbb{X}}|L'(R_L(x)'\tilde{\pi}_L)-L'(R_L(x)'\pi_L^*)|\frac{1}{\sqrt{N}}\sum_{i=1}^{N}|R_L(X_i)|=O_p\left(\zeta(L)^4L^{1/2}\right)$$

by lemma A.8, (v) and the Markov inequality. The second factor is $O(\sqrt{L})$, and the third, because

$$\mathbb{E}\left[\left|\frac{1}{N}\sum_{j=1}^{N}(W_j-e(X_j))R_L(X_j)\right|\right]\leq\frac{1}{N}\sqrt{\mathbb{E}\left[\left|\sum_{j=1}^{N}(W_j-e(X_j))R_L(X_j)\right|^2\right]}\leq$$

$$\frac{1}{\sqrt{N}}\sqrt{\mathbb{E}\left[(W_j-e(X_j)^2R_L(X_j)'R_L(X_j)\right]}=O(L^{1/2}N^{-1/2})$$

is by the Markov inequality $O_p(L^{1/2}N^{-1/2})$. Upon multiplication the first term is $O_p(\zeta(L)^4L^{3/2}N^{-1/2})$. The second term is a degenerate U-statistic

$$\left|\frac{1}{N\sqrt{N}}\sum_{i=1}^{N}\sum_{j=1}^{N}\frac{\mu(X_i)}{e_L^0(X_i)^2}(e(X_i)-W_i)L'(R_L(X_i)'\pi_L^*)R_L(X_i)'\Sigma_L^{-1}R_L(X_j)(W_j-e(X_j))\right|$$

We need to bound the variance of the kernel

$$\mathbb{E}\left[\left(\frac{\mu(X_i)}{e_L^0(X_i)^2}(e(X_i)-W_i)L'(R_L(X_i)'\pi_L^*)R_L(X_i)'\Sigma_L^{-1}R_L(X_j)(W_j-e(X_j))\right)^2\right]$$

$$=\mathbb{E}\left[\frac{\mu(X_i)^2}{e_L^0(X_i)^4}L'(R_L(X_i)'\pi_L^*)^2(e(X_i)-W_i)^2(W_j-e(X_j))^2(R_L(X_i)'\Sigma_L^{-1}R_L(X_j))^2\right]$$

$$\leq 16\cdot\sup_{x\in\mathbb{X}}\frac{\mu(x)^2}{e_L^0(x)^4}L'(R_L(x)'\pi_L^*)^2\mathbb{E}\left[(R_L(X_i)'\Sigma_L^{-1}R_L(X_j))^2\right]=O(L^2)$$

because $\mathbb{E}\left[(R_L(X_i)'\Sigma_L^{-1}R_L(X_j))^2\right]=\mathbb{E}\left[R_L(X_i)'\Sigma_L^{-1}R_L(X_j)R_L(X_j)'\Sigma_L^{-1}R_L(X_i)\right]=\operatorname{tr}\left((\Sigma_L^{-1})^2\right)$ Hence we conclude that (A.41) is $O_p\left(LN^{-1/2}\right)$ (see e.g. Van der Vaart (1998), Theorem 12.10). This ends the discussion of (A.24).

Expression (A.25) is bounded by

$$\left|\frac{1}{\sqrt{N}}\sum_{i=1}^N\frac{1}{e_L^0(X_i)}\left(\hat{\mu}_K(X_i)-\mu_K^0(X_i)\right)\left(\hat{e}_L(X_i)-W_i\right)\right|$$

$$\leq\left|\frac{1}{\sqrt{N}}\sum_{i=1}^N\frac{1}{e_L^0(X_i)}\left(\hat{\mu}_K(X_i)-\mu_K^0(X_i)\right)\left(\hat{e}_L(X_i)-e(X_i)\right)\right| \tag{A.43}$$

$$+\left|\frac{1}{\sqrt{N}}\sum_{i=1}^N\frac{1}{e_L^0(X_i)}\left(\hat{\mu}_K(X_i)-\mu_K^0(X_i)\right)\left(e(X_i)-W_i\right)\right| \tag{A.44}$$

For (A.43) we have the bound by lemma A.6(vi) and lemma A.8 (vii)

$$\left|\frac{1}{\sqrt{N}}\sum_{i=1}^N\frac{1}{e_L^0(X_i)}\left(\hat{\mu}_K(X_i)-\mu_K^0(X_i)\right)\left(\hat{e}_L(X_i)-e(X_i)\right)\right|\leq N^{1/2}\sup_{x\in\mathbb{X}}\frac{1}{e_L^0(x)}\sup_{x\in\mathbb{X}}\left|\hat{\mu}_K(x)-\mu_K^0(x)\right|\sup_{x\in\mathbb{X}}\left|\hat{e}_L(X_i)-e(X_i)\right|$$

$$=O_p\left(N^{1/2}\left(\zeta(K)^2K^{1/2}N^{-1/2}+\zeta(K)K^{-s_\mu/d}\right)\left(\zeta(L)L^{-s_e/(2d)}+\zeta(L)^2L^{1/2}N^{-1/2}\right)\right).$$

For (A.44) we have

$$\left|\frac{1}{\sqrt{N}}\sum_{i=1}^N\frac{1}{e_L^0(X_i)}\left(\hat{\mu}_K(X_i)-\mu_K^0(X_i)\right)\left(e(X_i)-W_i\right)\right|$$

$$\leq\left|\frac{1}{\sqrt{N}}\sum_{i=1}^N\frac{1}{e_L^0(X_i)}\left(\hat{\mu}_K(X_i)-\mu(X_i)\right)\left(e(X_i)-W_i\right)\right| \tag{A.45}$$

$$+\left|\frac{1}{\sqrt{N}}\sum_{i=1}^N\frac{1}{e_L^0(X_i)}\left(\mu(X_i)-\mu_K^0(X_i)\right)\left(e(X_i)-W_i\right)\right| \tag{A.46}$$

with (A.46) bounded by (lemma A.5)

$$\left|\frac{1}{\sqrt{N}}\sum_{i=1}^N\frac{1}{e_L^0(X_i)}\left(\mu(X_i)-\mu_K^0(X_i)\right)\left(e(X_i)-W_i\right)\right|$$

$$\leq N^{1/2}\sup_{x\in\mathbb{X}}\frac{1}{e_L^0(x)}\sup_{x\in\mathbb{X}}\left|\mu(x)-\mu_K^0(x)\right|\cdot 2=O_p\left(N^{1/2}K^{-s_\mu/d}\right)$$

which is a bound of type a.

To bound (A.45) we write

$$\left|\frac{1}{\sqrt{N}}\sum_{i=1}^{N}\frac{1}{e_L^0(X_i)}\left(\hat{\mu}_K(X_i)-\mu(X_i)\right)\left(e(X_i)-W_i\right)\right|$$

$$\leq\left|\frac{1}{\sqrt{N}}\sum_{i=1}^{N}\frac{1}{e_L^0(X_i)}\left(\hat{\mu}_K(X_i)-\mu_K^*(X_i)\right)\left(e(X_i)-W_i\right)\right| \tag{A.47}$$

$$+\left|\frac{1}{\sqrt{N}}\sum_{i=1}^{N}\frac{1}{e_L^0(X_i)}\left(\mu_K^*(X_i)-\mu(X_i)\right)\left(e(X_i)-W_i\right)\right| \tag{A.48}$$

Be lemma A.5 and A.6(ii) (A.48) is bounded by

$$N^{1/2}\sup_{x\in\mathbb{X}}\frac{1}{e_L^0(x)}\cdot\sup_{x\in\mathbb{X}}|\mu_K^*(x)-\mu(x)|\cdot 2$$

$$=O\left(N^{1/2}\zeta(K)K^{1/2-s_\mu/d}\right)$$

a type a bound.

For (A.47) the bound is

$$\left|\frac{1}{\sqrt{N}}\sum_{i=1}^{N}\frac{1}{e_L^0(X_i)}\left(\hat{\mu}_K(X_i)-\mu_K^*(X_i)\right)\left(e(X_i)-W_i\right)\right|$$

$$\leq\left|\frac{1}{\sqrt{N}}\sum_{i=1}^{N}\frac{1}{e_L^0(X_i)}R_K(X_i)'(\hat{\Omega}_{K,N_1}^{-1}-I_K)\left(\frac{1}{N_1}\sum_{j=1}^{N_1}R_K(X_j)Y_j\right)\left(e(X_i)-W_i\right)\right| \tag{A.49}$$

$$+\left|\frac{1}{\sqrt{N}}\sum_{i=1}^{N}\frac{1}{e_L^0(X_i)}R_K(X_i)'\left(\frac{1}{N_1}\sum_{j=1}^{N_1}R_K(X_j)Y_j-\mathbb{E}(R_K(X)\mu(X)|W=1)\right)\left(e(X_i)-W_i\right)\right| \tag{A.50}$$

with for (A.49) the type b bound

$$\left|\frac{1}{\sqrt{N}}\sum_{i=1}^{N}\frac{1}{e_L^0(X_i)}R_K(X_i)'(e(X_i)-W_i)(\hat{\Omega}_{K,N_1}^{-1}-I_K)\left(\frac{1}{N_1}\sum_{j=1}^{N_1}R_K(X_j)Y_j\right)\right|$$

$$\leq\left|\frac{1}{\sqrt{N}}\sum_{i=1}^{N}\frac{1}{e_L^0(X_i)}R_K(X_i)'(e(X_i)-W_i)\right|\left|\hat{\Omega}_{K,N_1}^{-1}-I_K\right|\left|\frac{1}{N_1}\sum_{j=1}^{N_1}R_K(X_j)Y_j\right|=O_p\left(\zeta(K)^2K^{3/2}N^{-1/2}\right)$$

Because the first factor is $O_p(\sqrt{K})$ by the Markov inequality, the third is $O_p(\zeta(K))$ by the Markov inequality and $\mathbb{E}(|R_K(X)Y|)=O(\zeta(K))$, and $\left|\hat{\Omega}_{K,N_1}^{-1}-I_K\right|=|\hat{\Omega}_{K,N_1}^{-1}|\left|\hat{\Omega}_{K,N_1}-I_K\right|=O_p(\sqrt{K})O_p(\zeta(K)K^{1/2}N^{-1/2})$. For (A.50)

$$\frac{1}{N\sqrt{N}}\frac{N}{N_1}\sum_{i=1}^{N}\sum_{j=1}^{N}\frac{1}{e_L^0(X_i)}R_K(X_i)'(e(X_i)-W_i)W_j\left(R_K(X_j)Y_j-\mathbb{E}(R_K(X)\mu(X)|W=1)\right)$$

is a degenerate U-statistic and the variance of the kernel is bounded by

$$4\cdot\sup_{x\in\mathbb{X}}\frac{1}{e_L^0(x)^2}\mathbb{E}\left[\left(R_K(X_i)'W_j\left(R_K(X_j)Y_j-\mathbb{E}(R_K(X)\mu(X)|W=1)\right)\right)^2\right]\leq 4\cdot\sup_{x\in\mathbb{X}}\frac{1}{e_L^0(x)^2}\mathbb{E}\left[R_K(X)'R_K(X)\right]$$

$$\cdot \mathbb{E}\left[W_j \left(R_K(X_j)'Y_j - \mathbb{E}(R_K(X)'\mu(X)|W=1)\right)\left(R_K(X_j)Y_j - \mathbb{E}(R_K(X)\mu(X)|W=1)\right)\right]$$

$$= 4 \cdot \sup_{x\in\mathbb{X}} \frac{1}{e_L^0(x)^2} \mathbb{E}\left[R_K(X)'R_K(X)\right] \cdot \left(\mathbb{E}\left[\mu_2(X)R_K(X)'R_K(X)|W=1\right] - \mathbb{E}(R_K(X)'\mu(X)|W=1)\mathbb{E}(R_K(X)\mu(X)|W=1)\right)$$

$$\leq 4 \cdot \sup_{x\in\mathbb{X}} \frac{1}{e_L^0(x)} \mathbb{E}\left[R_K(X)'R_K(X)\right] \cdot \mathbb{E}\left[\mu_2(X)R_K(X)'R_K(X)|W=1\right] = O(K^2)$$

with $\mu_2(x) = \mathbb{E}(Y^2|X=x)$ bounded on $\mathbb{X}$ and $\mathbb{E}\left[R_K(X)'R_K(X)\right] \leq \sup_{x\in\mathbb{X}} \frac{e}{e(x)} \mathbb{E}\left[R_K(X)'R_K(X)|W=1\right]$. Hence (A.50) is $O_p\left(KN^{-1/2}\right)$

Expression (A.26) is bounded by (lemma A.7)

$$\left|\frac{1}{\sqrt{N}} \sum_{i=1}^N \mu_K^0(X_i) \left(\frac{1}{e_L^0(X_i)} - \frac{1}{e(X_i)}\right) \left(\hat{e}_L(X_i) - W_i\right)\right|$$

$$\leq N^{1/2} \sup_{x\in\mathbb{X}} |\mu_K^0(x)| \cdot \sup_{x\in\mathbb{X}} \left|\frac{1}{e_L^0(x)} - \frac{1}{e(x)}\right| \cdot 2$$

$$= O_p\left(N^{1/2} L^{-s_e/(2d)}\right).$$

because by lemma A.5 $\mu_K^0(x)$ is bounded if K goes to ∞ with N.

Expression (A.27) is bounded by

$$\left|\frac{1}{\sqrt{N}} \sum_{i=1}^N \frac{1}{e(X_i)} \left(\mu_K^0(X_i) - \mu(X_i)\right) \cdot \left(\hat{e}_L(X_i) - W_i\right)\right|$$

$$\leq N^{1/2} \sup_{x\in\mathbb{X}} \frac{1}{e(x)} \cdot \sup_{x\in\mathbb{X}} \left|\mu_K^0(x) - \mu(x)\right| \cdot 2$$

$$= O_p\left(N^{1/2} L^{-s_\mu/d}\right).$$

Both bounds for (A.26) and (A.27) are of type a.

Finally, consider (A.28). Because $e(x)$ is bounded away from 0 on $\mathbb{X}$, $\mathbb{X}$ is a compact subset of $\mathbb{R}^d$, and $\mu(x)$ and $e(x)$ are s_μ and s_e times continuously differentiable, it follows that $\frac{\mu(X)}{e(X)}$ is $\min(s_\mu, s_e)$ times continuously differentiable. By lemma A.5 there is δ_L^0 such that

$$\sup_{x\in\mathbb{X}} \left|\frac{\mu(x)}{e(x)} - R_L(x)'\delta_L^0\right| = O\left(L^{-\frac{\min(s_\mu, s_e)}{d}}\right). \tag{A.51}$$

Hence,

$$\left|\frac{1}{\sqrt{N}} \sum_{i=1}^N \frac{\mu(X_i)}{e(X_i)} (\hat{e}_L(X_i) - W_i)\right| \tag{A.52}$$

$$\leq \left|\frac{1}{\sqrt{N}} \sum_{i=1}^N \left(\frac{\mu(X_i)}{e(X_i)} - R_L(X_i)'\delta_L^0\right) \left(\hat{e}_L(X_i) - W_i\right)\right| + \left|\frac{1}{\sqrt{N}} \sum_{i=1}^N R_L(X_i)'\delta_L^0 (\hat{e}_L(X_i) - W_i)\right|$$

$$= \left|\frac{1}{\sqrt{N}} \sum_{i=1}^N \left(\frac{\mu(X_i)}{e(X_i)} - R_L(X_i)'\delta_L^0\right) \left(\hat{e}_L(X_i) - W_i\right)\right|, \tag{A.53}$$

because the second term vanishes as a result of the first order conditions for the series logit estimator which imply that $\sum_i R_L(X_i)(\hat{e}(X_i) - W_i) = 0$. The last expression, (A.53) can be bounded by

$$\sup_{x\in\mathbb{X}}\left|\frac{\mu(x)}{e(x)} - R_L(x)'\delta_L^0\right| \frac{1}{\sqrt{N}}\sum_{i=1}^N |\hat{e}(X_i) - W_i| \leq 2\cdot\sqrt{N}\sup_{x\in\mathbb{X}}\left|\frac{\mu(x)}{e(x)} - R_L(x)'\delta_L^0\right| = O\left(N^{1/2}L^{-\frac{\min(s_\mu,s_e)}{d}}\right),$$

Combining the bounds finishes the proof of the first assertion in theorem 3.1. We find that the bound is

$$O_p\left(N^{-1/2}\zeta(K)^2\zeta(L)K^{1/2}L^{1/2}\right) + O_p\left(N^{-1/2}\zeta(K)^3K^{1/2}\right) + O_p\left(N^{-1/2}\zeta(L)^5L\right) \tag{A.54}$$

$$+O_p\left(N^{1/2}\zeta(K)^2\zeta(L)^2K^{-s_\mu/d}L^{-s_e/(2d)}\right) + O_p\left(\zeta(K)^2\zeta(L)L^{1/2}K^{-s_\mu/d}\right) + O_p\left(N^{1/2}\zeta(K)K^{-s_\mu/d}\right)$$

$$+O_p\left(\zeta(L)^4L^{1/2}L^{-s_e/(2d)}\right)+O_p\left(N^{1/2}\zeta(L)^2L^{-s_e/d}\right)+O_p\left(N^{1/2}\zeta(L)L^{-s_e/(2d)}\right)+O_p\left(\zeta(K)^2\zeta(L)^2K^{1/2}L^{-s_e/(2d)}\right)$$

Next, consider part (ii) of theorem 3.1. By the triangle inequality we have because $\frac{1}{\hat{e}_L(X)} = \frac{1}{e(X)} - \frac{\hat{e}_L(X)-e(X)}{\hat{e}_L(X)e(X)}$

$$\left|\sqrt{N}\cdot(\hat{\tau}_{mod} - \hat{\tau}_{hir})\right| = \left|\frac{1}{\sqrt{N}}\sum_{i=1}^N \frac{W_i(Y_i - \hat{\mu}_K(X_i))}{\hat{e}_L(X_i)}\right|$$

$$\leq \left|\frac{1}{\sqrt{N}}\sum_{i=1}^N \frac{W_i(Y_i - \hat{\mu}_K(X_i))}{e(X_i)}\right| \tag{A.55}$$

$$+\left|\frac{1}{\sqrt{N}}\sum_{i=1}^N \frac{W_i}{\hat{e}_L(X_i)e(X_i)}(\hat{e}_L(X_i) - e(X_i))(Y_i - \hat{\mu}_K(X_i))\right| \tag{A.56}$$

First consider (A.56) that we bound by

$$\left|\frac{1}{\sqrt{N}}\sum_{i=1}^N \frac{W_i}{\hat{e}_L(X_i)e(X_i)}(\hat{e}_L(X_i) - e(X_i))(Y_i - \mu(X_i))\right| \tag{A.57}$$

$$+\left|\frac{1}{\sqrt{N}}\sum_{i=1}^N \frac{W_i}{\hat{e}_L(X_i)e(X_i)}(\hat{e}_L(X_i) - e(X_i))(\hat{\mu}_K(X_i) - \mu(X_i))\right| \tag{A.58}$$

Note that (A.58) can be bounded in the same way as (A.43) using lemma A.8 (vii) and lemma A.6 (vi), so that this term is $O_p\left(N^{1/2}\left(\zeta(K)^2K^{1/2}N^{-1/2} + \zeta(K)K^{-s_\mu/d}\right)\left(\zeta(L)^2L^{-s_e/(2d)} + \zeta(L)L^{1/2}N^{-1/2}\right)\right)$. Also note that (A.57) can be bounded by

$$\left|\frac{1}{\sqrt{N}}\sum_{i=1}^N \frac{W_i}{\hat{e}_L(X_i)e(X_i)}(\hat{e}_L(X_i) - e_L^*(X_i))(Y_i - \mu(X_i))\right| \tag{A.59}$$

$$+\left|\frac{1}{\sqrt{N}}\sum_{i=1}^N \frac{W_i}{\hat{e}_L(X_i)e(X_i)}(e_L^*(X_i) - e(X_i))(Y_i - \mu(X_i))\right| \tag{A.60}$$

We have that (A.59) can be bounded as (A.34) so that by the bounds on (A.37), (A.39),(A.42), and (A.41), we find the bound $O_p\left(N^{1/2}\zeta(L)^3L^{1/2}N^{-1/2}(\zeta(L)^2L^{1/2}N^{-1/2} + \zeta(L)L^{-s_e/(2d)})\right)+O_p\left(\zeta(L)^3L^{1/2}N^{-1/2}\right)+O_p\left(LN^{-1/2}\right) + O_p\left(L^{-s_e/(2d)}\zeta(L)^2\right)$. For (A.60) we have the bound (lemma A.7 and A.8(ii))

$$\sqrt{N}\sup_{x\in\mathbb{X}}\frac{1}{\hat{e}_L(x)e(x)}\sup_{x\in\mathbb{X}}|e_L^*(x) - e(x)|\frac{1}{N}\sum_{i=1}^N W_i\left|(Y_i - \mu(X_i))\right| = O_p\left(N^{1/2}\zeta(L)L^{-s_e/(2d)}\right)$$

Finally we consider (A.55). Because $e(x)$ is s times continuously differentiable and bounded away from zero, by lemma A.5 there is a δ_K^0 such that

$$\sup_{x\in\mathbb{X}}\left|\frac{1}{e(x)} - R_K(x)'\delta_K^0\right| = O\left(K^{-s_e/d}\right).$$

¿From the normal equations of the regression of Y on $R_K(X)$ it follows that

$$\sum_{i=1}^N W_i(Y_i - \hat{\mu}_K(X_i))R_K(X_i) = 0.$$

Hence

$$\begin{aligned}
&\left|\frac{1}{\sqrt{N}}\sum_{i=1}^N \frac{W_i(Y_i - \hat{\mu}_K(X_i))}{e(X_i)}\right| \\
&\quad\le \left|\frac{1}{\sqrt{N}}\sum_{i=1}^N W_i(Y_i - \hat{\mu}_K(X_i))\cdot\left(\frac{1}{e(X_i)} - R_K(X_i)\delta_K^0\right)\right| \\
&\qquad + \left|\frac{1}{\sqrt{N}}\sum_{i=1}^N W_i(Y_i - \hat{\mu}_K(X_i))\cdot R_K(X_i)\delta_K^0\right| \\
&\quad\le \left|\frac{1}{\sqrt{N}}\sum_{i=1}^N W_i(Y_i - \mu(X_i))\cdot\left(\frac{1}{e(X_i)} - R_K(X_i)\delta_K^0\right)\right| \\
&\qquad + \left|\frac{1}{\sqrt{N}}\sum_{i=1}^N W_i(\mu(X_i) - \hat{\mu}_K(X_i))\cdot\left(\frac{1}{e(X_i)} - R_K(X_i)\delta_K^0\right)\right| \\
&\quad\le N^{1/2}\frac{1}{N}\sum_{i=1}^N |W_i(Y_i - \mu(X_i))|\cdot\sup_{x\in\mathbb{X}}\left|\frac{1}{e(x)} - R_K(x)\delta_K^0\right| \\
&\qquad + N^{1/2}\sup_{x\in\mathbb{X}}|\mu(x) - \hat{\mu}_K(x)|\cdot\sup_{x\in\mathbb{X}}\left|\frac{1}{e(x)} - R_K(x)\delta_K^0\right| \\
&\quad= O_p(N^{1/2}K^{-s_e/d}) + O_p(N^{1/2}K^{-s_e/d}(\zeta(K)^2K^{1/2}N^{-1/2} + \zeta(K)K^{-s_\mu/d}))
\end{aligned}$$

by lemma A.6(vi). Combing the bounds we find that the upper bound is

$$\begin{aligned}
&O_p\left(N^{-1/2}\zeta(K)^2\zeta(L)K^{1/2}L^{1/2}\right) + O_p\left(N^{-1/2}\zeta(L)^5L\right) \\
&+O_p\left(N^{1/2}\zeta(K)^2\zeta(L)^2K^{-s_\mu/d}L^{-s_e/(2d)}\right) + O_p\left(\zeta(K)\zeta(L)L^{1/2}K^{-s_\mu/d}\right) \\
&+O_p\left(\zeta(L)^4L^{1/2}L^{-s_e/(2d)}\right) + O_p\left(\zeta(K)^2\zeta(L)^2K^{1/2}L^{-s_e/(2d)}\right) \\
&+O_p\left(N^{1/2}\zeta(K)^2K^{1/2}K^{-s_e/d}\right) + O_p\left(N^{1/2}\zeta(K)K^{1/2}K^{-s_e/d-s_\mu/d}\right)
\end{aligned} \tag{A.61}$$

Finally, consider part (*iii*) of the theorem. First we derive a bound on

$$\left|N^{1/2}\left(\sum_{i=1}^N \frac{W_i}{\hat{e}_L(X_i)}\Big/ N - 1\right)\right|. \tag{A.62}$$

We use $\frac{1}{\hat{e}_L(X)} = \frac{1}{e(X_i)} - \frac{1}{e(X)\hat{e}_L(X)}(\hat{e}_L(X) - e(X))$ to obtain

$$\left| N^{1/2} \left(\sum_{i=1}^{N} \frac{W_i}{\hat{e}_L(X_i)} \Big/ N - 1 \right) \right| = \left| \frac{1}{\sqrt{N}} \sum_{i=1}^{N} \frac{W_i - \hat{e}_L(X_i)}{\hat{e}_L(X_i)} \right|$$

$$\leq \left| \frac{1}{\sqrt{N}} \sum_{i=1}^{N} \frac{W_i - \hat{e}_L(X_i)}{e(X_i)} \right| \tag{A.63}$$

$$+ \left| \frac{1}{\sqrt{N}} \sum_{i=1}^{N} \frac{W_i - \hat{e}_L(X_i)}{e(X_i)\hat{e}_L(X_i)} (\hat{e}_L(X_i) - e(X_i)) \right| \tag{A.64}$$

First consider (A.64) that we bound by

$$\left| \frac{1}{\sqrt{N}} \sum_{i=1}^{N} \frac{W_i - \hat{e}_L(X_i)}{e(X_i)\hat{e}_L(X_i)} (\hat{e}_L(X_i) - e(X_i)) \right|$$

$$\leq \left| \frac{1}{\sqrt{N}} \sum_{i=1}^{N} \frac{W_i - e(X_i)}{e(X_i)\hat{e}_L(X_i)} (\hat{e}_L(X_i) - e(X_i)) \right| \tag{A.65}$$

$$+ \left| \frac{1}{\sqrt{N}} \sum_{i=1}^{N} \frac{(\hat{e}_L(X_i) - e(X_i))^2}{e(X_i)\hat{e}_L(X_i)} \right| \tag{A.66}$$

Note that (A.65) can be bounded as (A.33) with the only differend that e_L^0 must be replaced with e. Hence from (A.35), (A.37),(A.39),(A.42), and (A.41) we find the bound $O_p\left(N^{1/2}\zeta(L)L^{-s_e/(2d)}\right) + O_p\left(N^{1/2}\zeta(L)^3L^{1/2}N^{-1/2}(\zeta(L)^2L^{1/2}N^{-1/2} + \zeta(L)L^{-s_e/(2d)})\right) + O_p\left(\zeta(L)^3L^{1/2}N^{-1/2}\right) + O_p\left(\zeta(L)^2L^{1/2}L^{-s_e/(2d)}\right) + O_p\left(LN^{-1/2}\right)$. For (A.66) we note the similarity with (A.32) (again e_L^0 must be replaced with e), so that we have the bound $O_p\left(N^{1/2}(\zeta(L)^2L^{1/2}N^{-1/2} + \zeta(L)L^{-s_e/(2d)})^2\right)$. Finally consider (A.63). Because $e(x)$ is bounded from 0 on $\mathbb{X}$) and is s_e times continuously differentiable, there is a sequence of δ_L such that for some finite C we have

$$\sup_{x\in\mathbb{X}} \left| \frac{1}{e(x)} - R_L(x)'\delta_L \right| = O(L^{-s_e/d}).$$

Then

$$\left| \frac{1}{\sqrt{N}} \sum_{i=1}^{N} \frac{W_i - \hat{e}_L(X_i)}{e(X_i)} \right|$$

$$\leq \left| \frac{1}{\sqrt{N}} \sum_{i=1}^{N} (W_i - \hat{e}_L(X_i))\, R_L(X_i)'\delta_L \right| + \left| \frac{1}{\sqrt{N}} \sum_{i=1}^{N} (W_i - \hat{e}_L(X_i)) \cdot \left(R_L(X_i)'\delta_L - \frac{1}{e(X_i)} \right) \right|.$$

Because of the first order conditions for $\hat{\pi}_L$ the first term vanishes. The second term is bounded by

$$2 \cdot N^{1/2} \cdot \sup_{x\in\mathbb{X}} \left| R_L(x)'\delta_L - \frac{1}{e(x)} \right| = O(N^{1/2}L^{-s_e/d}).$$

Next, define

$$C_N = \frac{1}{N} \sum_{i=1}^{N} \frac{W_i \cdot Y_i}{\hat{e}_L(X_i)} - 1,$$

so that

$$\hat{\beta}_{hir,1} - \hat{\beta}_{hir,2} = \frac{1}{N}\sum_{i=1}^{N} \frac{W_i \cdot Y_i}{\hat{e}_L(X_i)} \cdot \left(1 - \frac{1}{1+C_N}\right).$$

Hence

$$\begin{aligned}
&\sqrt{N} \cdot \left|\hat{\beta}_{hir,1} - \hat{\beta}_{hir,2}\right| \\
&\qquad \leq \left|\frac{1}{N}\sum_{i=1}^{N} \frac{W_i \cdot Y_i}{\hat{e}_L(X_i)}\right| \cdot \left|\frac{\sqrt{N} \cdot C_N}{1+C_N}\right|.
\end{aligned}$$

The first factor is $\hat{\tau}_{hir}$ and is $O_p(1)$ if the rate restrictions that ensure weak consistency are satisfied. The second term is bounded by bound on $\sqrt{N}C_N$ derived above and this is also the bound on the difference, so that the bound is

$$O_p\left(N^{-1/2}\zeta(L)^5 L\right) + O_p\left(\zeta(L)^4 L^{1/2} L^{-s_e/(2d)}\right) + O_p\left(N^{1/2}\zeta(L) L^{-s_e/(2d)}\right) + O_p\left(N^{1/2}\zeta(L)^2 L^{-s_e/d}\right) \tag{A.67}$$

□

Lemma A.9 *If*

$$\sup_{K \in \mathcal{K}_N} \frac{|E_N(K) - C_N(K)|}{E_N(K)} \xrightarrow{p} 0$$

then

$$\frac{E_N(\hat{K})}{\inf_{K \in \mathcal{K}_N} E_N(K)} \xrightarrow{p} 1$$

Proof: Define

$$\tilde{K} = \text{argmin}_{K \in \mathcal{K}_N} E_N(K)$$

For N large enough we have that for any $\delta, \eta > 0$

$$\Pr\left(\sup_{K \in \mathcal{K}_N} \left|\frac{C_N(K)}{E_N(K)} - 1\right| < \delta\right) > 1 - \eta$$

Hence if N is sufficiently large then with probability of at least $1 - \eta$

$$\frac{1+\delta}{1-\delta} \geq \frac{C_N(\tilde{K})}{(1-\delta)E_N(\tilde{K})} \geq \frac{C_N(\hat{K})}{(1-\delta)E_N(\tilde{K})} \geq \frac{E_N(\hat{K})}{E_N(\tilde{K})} \geq 1$$

and the conclusion follows because δ, η are arbitrary. □

Proof of Theorem 4.1: We have

$$C_N(K) = E_N(K) + \frac{2}{N} u'_{N_1} A_{K,N_1} a_{N_1} a'_{N_1} A_{K,N_1} \mu_{N_1} + \frac{1}{N}(a'_{N_1} A_{K,N_1} u_{N_1} u'_{N_1} A_{K,N_1} a_{N_1} - \sigma^2 a'_{N_1} A_{K,N_1} a_{N_1})$$

We need to show

$$\sup_{K \in \mathcal{K}_N} \frac{\frac{1}{N}\left|u'_{N_1} A_{K,N_1} a_{N_1} a'_{N_1} A_{K,N_1} \mu_{N_1}\right|}{E_N(K)} \xrightarrow{p} 0 \tag{A.68}$$

and

$$\sup_{K\in\mathcal{K}_N} \frac{\frac{1}{N}\left|a'_{N_1}A_{K,N_1}u_{N_1}u'_{N_1}A_{K,N_1}a_{N_1} - \sigma^2 a'_{N_1}A_{K,N_1}a_{N_1}\right|}{E_N(K)} \xrightarrow{p} 0 \tag{A.69}$$

We consider (A.68) first.

We have for $\delta > 0$ (using the Markov inequality for the final bound)

$$\Pr\left(\sup_{K\in\mathcal{K}_N}\frac{1}{N}\frac{|a'_{N_1}A_{K,N_1}\mu_{N_1}||u'_{N_1}A_{K,N_1}a_{N_1}|}{E_N(K)} > \delta\right) \leq \sum_{K\in\mathcal{K}_N}\Pr\left(\frac{1}{N}\frac{|a'_{N_1}A_{K,N_1}\mu_{N_1}||u'_{N_1}A_{K,N_1}a_{N_1}|}{E_N(K)} > \delta\right)$$

$$\leq \sum_{K\in\mathcal{K}_N}\frac{|a'_{N_1}A_{K,N_1}\mu_{N_1}|^m E(|u'_{N_1}A_{K,N_1}a_{N_1}|^m)}{\delta^m N^m E_N(K)^m}$$

By Theorem 2 in Whittle (1960) and assumption 4.1

$$E(|u'_{N_1}A_{K,N_1}a_{N_1}|^m) \leq C(a'_{N_1}A_{K,N_1}a_{N_1})^{\frac{m}{2}}$$

We have by lemmas A.7, A.8(ii) and assumption 3.2

$$a'_{N_1}A_{K,N_1}a_{N_1} \leq CN\zeta(K)^2K^{-s_e/d} \tag{A.70}$$

Also

$$NE_N(K) \geq (a'_{N_1}A_{K,N_1}\mu_{N_1})^2 \tag{A.71}$$

so that

$$\sum_{K\in\mathcal{K}_N}\frac{|a'_{N_1}A_{K,N_1}\mu_{N_1}|^m E(|u'_{N_1}A_{K,N_1}a_{N_1}|^m)}{\delta^m N^m E_N(K)^m} \leq \sum_{K\in\mathcal{K}_N} C\frac{\zeta(K)^m K^{-ms_e/(2d)}}{E_N(K)^{\frac{m}{2}}}.$$

Hence sufficient for (A.68) is that:

$$\sum_{K\in\mathcal{K}_N}\frac{\zeta(K)^m K^{-ms_e/(2d)}}{E_N(K)^{\frac{m}{2}}} \to 0$$

We have

$$\sum_{K\in\mathcal{K}_N}\frac{\zeta(K)^m K^{-ms_e/(2d)}}{E_N(K)^{\frac{m}{2}}} \leq \frac{1}{\inf_{K\in\mathcal{K}_N}E_N(K)^{\frac{m}{2}}}\sum_{K\in\mathcal{K}_N}\zeta(K)^m K^{-ms_e/(2d)}$$

With $\mathcal{K}_N = \{N^{\nu_0},\ldots,N^{\nu_1}\}$, this holds if

$$N^{-\nu_1+\nu_0 m(s_e/(2d)-1)}\inf_{K\in\mathcal{K}_N}E_N(K)^{\frac{m}{2}} \to \infty \tag{A.72}$$

We now consider (A.69). We have by Theorem 2 in Whittle (1960) and assumption 4.1

$$\Pr\left(\sup_{K\in\mathcal{K}_N}\frac{\left|a'_{N_1}A_{K,N_1}u_{N_1}u'_{N_1}A_{K,N_1}a_{N_1} - \sigma^2 a'_{N_1}A_{K,N_1}a_{N_1}\right|}{NE_N(K)} > \delta\right) \leq$$

$$\leq \sum_{K\in\mathcal{K}_N}\Pr\left(\frac{\left|a'_{N_1}A_{K,N_1}u_{N_1}u'_{N_1}A_{K,N_1}a_{N_1} - \sigma^2 a'_{N_1}A_{K,N_1}a_{N_1}\right|}{NE_N(K)} > \delta\right) \leq$$

$$\leq \sum_{K\in\mathcal{K}_N} \frac{E\left(\left|a'_{N_1}A_{K,N_1}u_{N_1}u'_{N_1}A_{K,N_1}a_{N_1} - \sigma^2 a'_{N_1}A_{K,N_1}a_{N_1}\right|^m\right)}{\delta^m N^m E_N(K)^m} \leq$$

$$\leq \sum_{K\in\mathcal{K}_N} \frac{C\mathrm{tr}\left((A_{K,N_1}a_{N_1}a'_{N_1}A_{K,N_1})^2\right)^{\frac{m}{2}}}{\delta^m N^m E_N(K)^m} \leq C\sum_{K\in\mathcal{K}_N} \frac{(a'_{N_1}A_{K,N_1}a_{N_1})^m}{\delta^m N^m E_N(K)^m}$$

By (A.70)

$$\sum_{K\in\mathcal{K}_N} \frac{(a'_{N_1}A_{K,N_1}a_{N_1})^m}{\delta^m N^m E_N(K)^m} \leq C\sum_{K\in\mathcal{K}_N} \frac{\zeta(K)^{2m}K^{-ms_e/d}}{E_N(K)^m}$$

so that we need

$$\sum_{K\in\mathcal{K}_N} \frac{\zeta(K)^{2m}K^{-ms_e/d}}{E_N(K)^m} \to 0$$

Under the earlier assumptions on $\zeta(K)$ and the dimension of $\mathcal{K}_N$ we need

$$N^{\kappa(m(s_e/d-2)-1)} \inf_{K\in\mathcal{K}_N} E_N(K)^m \to \infty$$

which is implied by (A.72). □

REFERENCES

Abadie, A., and G. Imbens, (2005), "Large Sample Properties of Matching Estimators for Average Treatment Effects," *Econometrica*, forthcoming.

Andrews, D. (1991): "Asymptotic Optimality of Generalized C_L, Cross-validation, and Generalized Cross-validation in Regression with Heteroskedastic Errors," *Journal of Econometrics*, 47, 359-377.

Barnow, B.S., G.G. Cain and A.S. Goldberger (1980), "Issues in the Analysis of Selectivity Bias," in *Evaluation Studies* , vol. 5, ed. by E. Stromsdorfer and G. Farkas. San Francisco: Sage.

Chen, X., Hong, H., and Tarozzi, A., (2004), " Semiparametric Efficiency in GMM Models of Nonclassical Measurement Error, Missing Data and Treatment Effects. " Working Paper.

Crump, R., V. J. Hotz, G. Imbens and O. Mitnik, (2005), "Moving the Goalposts: Addressing Limited Overlap in Estimation of Average Treatment Effects by Changing the Estimand," unpublished manuscript, Department of Economics, UC Berkeley.

Dehejia, R., and S. Wahba, (1999), "Causal Effects in Nonexperimental Studies: Reevaluating the Evaluation of Training Programs", *Journal of the American Statistical Association*, 94: 1053-1062.

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., and R. Robb, (1984), "Alternative Methods for Evaluating the Impact of Interventions," in Heckman and Singer (eds.), *Longitudinal Analysis of Labor Market Data*, Cambridge, Cambridge University Press.

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

Heckman, J., H. Ichimura, J. Smith, and P. Todd, (1998), "Characterizing Selection Bias Using Experimental Data," *Econometrica* 66, 1017-1098.

Hirano, K., G. Imbens, and G. Ridder (2003), "Efficient Estimation of Average Treatment Effects Using the Estimated Propensity Score. " *Econometrica*, 71(4): 1161-1189.

Holland, P. (1986), "Statistics and Causal Inference," *Journal of the American Statistical Association*, 81, 945-970.

Ichimura, H., and O. Linton, (2001), Asymptotic Expansions for some Semiparametric Program Evaluation Estimators." Institute for Fiscal Studies, cemmap working paper cwp04/01.

Imbens, G., (2004), "Nonparametric Estimation of Average Treatment Effects Under Exogeneity: A Review, *Review of Economics and Statistics*, 86(1): 1-29.

Li, K. (1987), "Asymptotic Optimality for C_p, C_L, Cross-validation and Generalized Cross-validation: Discrete Index Set," *Annals of Statistics*, 15(3), 958-975.

Newey, W., (1994), "The Asymptotic Variance of Semiparametric Estimators," *Econometrica* 62, 1349-1382.

Newey, W. (1995), "Convergence Rates for Series Estimators," in *Advances in Econometrics and Quantitative Economics: Essays in Honor of Professor C. R. Rao*, Maddala, Phillips, and Srinivasan (eds.), Cambridge, Basil Blackwell.

Newey, W., and D. McFadden (1994), "Large Sample Estimation," in *Handbook of Econometrics*, Vol. 4, Engle and McFadden (eds.), North Holland.

Robins, J., A. Rotnitzky, and L. Zhao (1995), "Analysis of Semiparametric Regression Models for Repeated Outcomes in the Presence of Missing Data," *Journal of the American Statistical Association*, 90 (429), 106-121.

Rosenbaum, P., and D. Rubin, (1983a), "The Central Role of the Propensity Score in Observational Studies for Causal Effects", *Biometrika*, 70, 41-55.

Rotnitzky, A., and J. Robins (1995), "Semiparametric Regression Estimation in the Presence of Dependent Censoring," *Biometrika* 82 (4), 805-820.

Rubin, D. (1974), "Estimating Causal Effects of Treatments in Randomized and Non-randomized Studies," *Journal of Educational Psychology*, 66, 688-701.

Rubin, D. B., (1978), "Bayesian inference for causal effects: The Role of Randomization", *A*nnals of Statistics, 6:34–58.