UNIVERSITY OF CALIFORNIA

Los Angeles

Higher-Order Optimal Estimation of Binary Average Treatment Effects

A dissertation submitted in partial satisfaction of the requirements for the degree Doctor of Philosophy in Economics

by

Paul Joseph Gift

2002

ProQuest Information and Learning Company
300 North Zeeb Road
P.O. Box 1346
Ann Arbor, MI 48106-1346

The dissertation of Paul Joseph Gift is approved.

Keisuke Hirano

V. Joseph Hotz

Geert Ridder

Guido Imbens, Committee Chair

University of California, Los Angeles

2002

To my dad, Michael J. Gift, Ph.D.,

I still remember what you said to me one day after coming home from playing basketball as a kid in Indiana. You asked me what I wanted to do when I grew up. I said I wanted to be a professional basketball player. You encouraged those dreams but also instilled in me the importance of an education as a backup in case those dreams didn't come true. Well, they didn't. Thank you so much for helping me all along the way, urging me to find something I like and to excel in it.

TABLE OF CONTENTS

FIGURES AND TABLES

Figures

Tables

ACKNOWLEDGEMENTS

I'd like to thank Guido Imbens for extremely helpful guidance and support throughout this entire process, even while out of town. Your instruction and research support and tips have been incredibly valuable throughout my coursework phase and dissertation phase here at UCLA. I really appreciate all the time and availability you've provided me the past few years. I would also like to give a thank you to Keisuke Hirano. I learned a lot from working with you on the GPS stuff and thank you for all of your time and advice on my dissertation. I'd like to thank Joe Hotz, Joris Pinkse, and Geert Ridder for valuable comments, meetings, and advice. Thank you to all my friends and fellow economics graduate students who made my time here at UCLA very interesting and a lot of fun. Finally, Thank you to my roommates, Mike Backstrom, Franz Bautista, and Jason Yanagihara for not making too much noise when I had to work. Oh, and thank you to Q's for being a fun bar across the street at which to relax, listen to good music, and beat my aforementioned roommates in pool.

Lastly, I'd like to send a very special thank you to my family, my dad Michael, my mom Cynthia, and my brother Dane. Except when Dane would pick on me when we were younger, you have all been extremely supportive of all that I do and it means a lot to me. Thank you very much.

VITA

September 8, 1976	Born, Florida, USA
1997	B.A. Economics Pepperdine University Malibu, California
1998-99	Teaching Assistant Department of Economics University of California, Los Angeles
2000	M.A. Economics University of California, Los Angeles Los Angeles, California
2000	Adjunct Professor Business Division Pepperdine University
2001	Teaching Assistant Department of Economics University of California, Los Angeles
2001	Adjunct Professor Graziadio School of Business and Management Pepperdine University
2002	Adjunct Professor Business Division Pepperdine University
2002 –	Economist Economic Analysis, LLC Century City, California

ABSTRACT OF THE DISSERTATION

Higher-Order Optimal Estimation of Binary Average Treatment Effects

by

Paul Joseph Gift
Doctor of Philosophy in Economics
University of California, Los Angeles, 2002
Professor Guido Imbens, Chair

Many times economists and other scientists are interested in estimating the causal effect of a binary treatment on a scalar outcome. In this paper, I propose a procedure for higher-order optimal estimation of average treatment effects with finite sample sizes. Hahn (1998) shows that adjustment for the known propensity score is in general inefficient. Hirano, Imbens, and Ridder (2000) show that adjusting for the non-parametrically estimated propensity score by weighting is consistent and asymptotically efficient. However, this estimator will tend not to be the optimal estimator in small samples. I propose estimation of average treatment effects within a GMM system of moments. A finite number of auxiliary moment restrictions reflecting one's knowledge of the propensity score as the conditional expectation of the treatment indicator can be added to increase efficiency. Asymptotic expansions are used to derive a higher-order mean squared error approximation for the resulting estimator. Following the work of Donald and Newey (1999), I develop a higher-order asymptotically optimal criterion for the selection of auxiliary moments. In an experimental setting, Monte Carlo simulations show that this moment selection procedure performs very well under a variety of data generating processes. In many cases, a significant efficiency gain obtains relative to the true propensity score estimator and an estimator similar to the HIR non-parametric propensity score estimator. In a non-experimental setting, the GMM procedure above is adjusted to account for the fact that the propensity score is unknown. The propensity score is estimated in a flexible Logit framework. This Unknown Propensity Score GMM estimator is more efficient than the HIR estimator in finite samples, with the efficiency gain going to zero as sample size increases. However, the non-experimental moment selection criterion is not asymptotically optimal and performs poorly in small samples. Finally, through simulations, I develop a finite sample compliment to Hahn (1998). Specifically, the propensity score is *not* ancillary for estimation of average treatment effects, relative to the HIR estimator and the Unknown Propensity Score GMM estimator. These results hold whether the econometrician has data on a vector of covariates or just a scalar.

1. Chapter 1

Higher-Order Optimal Estimation of Binary Average Treatment Effects with Experimental Data

1.1 Introduction

Many times researchers in economics and other sciences are interested in estimating the effect of a binary treatment on a scalar outcome. Some examples are the effect of job training on future earnings, the effect of a takeover on firm productivity, or the effect of a new drug or vaccine on life span. This is generally done in one of two settings, experimental and non-experimental. In an experimental setting, selection into treatment status is undertaken randomly by the experiment coordinators. Thus, treatment status is independent of potential outcomes. This type of data can be very useful for estimating average treatment effects. A simple, common estimator will yield unbiased, but possibly sub-optimal, results. Employing data on pre-treatment covariates can lead to significant efficiency gains in estimation in this setting. However, other times one is left with no other viable alternative but to use data from non-experimental studies. In this situation, economists typically believe that assignment to treatment is not independent of potential outcomes. In other words, the economics agents involved have self-selected into a given treatment status. If one has data on pre-treatment variables in which self-selection is based, one can assume unconfoundedness. This implies that assignment to treatment is independent of potential outcomes conditional on the pre-treatment variables. Controlling for these pre-treatment variables will remove all biases associated with simple estimation of the average treatment effect.

Rosenbaum and Rubin (1983) show that all biases can also be removed by conditioning on the scalar propensity score rather than conditioning on the entire (possible) vector of pre-treatment variables. Hirano, Imbens, and Ridder (2000) (henceforth HIR) show that weighting by the inverse of a non-parametric estimate of the propensity score leads to consistent and asymptotically efficient average treatment effect estimates. They show that the normalized average treatment effect estimator converges in distribution to a normal random variable with mean zero and the semi-parametric efficiency bound as its variance. They compare this to the estimator obtained by weighting by the inverse of the true propensity score and note that the true propensity score estimator is unbiased but less efficient than the HIR non-parametric propensity score estimator. However, all of HIR's calculations are asymptotic ones. In finite samples, the semi-parametric efficiency bound cannot be reached. This chapter will show that in small samples there can be an efficiency gain to constructing an estimator that efficiently incorporates some of the econometrician's auxiliary information about the propensity score. This holds true in *both* experimental and non-experimental studies. Higher-order optimality is considered within a given class of estimators. This class will be those estimators obtained from a particular set of GMM moment restrictions. The estimator imposes a finite amount of additional information from a subset of the pre-treatment variables. This is to be contrasted with the true propensity score estimator, which contains no additional information and the HIR non-parametric propensity score estimator, which contains an infinite amount of additional information. Optimality in this case is defined as minimum mean squared error (MSE).

In what follows, this chapter examines estimation of an average treatment effect in the context of a GMM system of moments. The first moment by itself leads to the estimator of the average treatment effect when weighting is done with the true propensity score. Subsequent moments include additional information stemming from knowledge of the propensity score. The question of interest is: Given finite sample size, how many auxiliary moments should be included in order to minimize the MSE of the resulting estimator? When sample size is large, if all possible additional moments are included, the resulting estimator converges in distribution to the HIR non-parametric propensity score estimator. Since the HIR estimator is efficient, the econometrician should include all additional moments (e.g. use the HIR estimator). However, when sample size is small, moment selection becomes an important practical consideration in any particular estimation procedure. In this setting, there exists a trade off between the benefit of the information of a marginal moment restriction and the cost of the additional noise created by that marginal moment restriction. This is analogous to the MSE test in OLS regression. In that setting, one wishes to determine if inclusion of an omitted variable lowers the MSE of a parameter of interest. In other words, one wishes to determine whether the benefit of inclusion of a marginal variable in terms of lower bias outweighs the cost of inclusion of that marginal variable in terms of increased variance.

A higher-order asymptotically optimal moment selection criterion function is derived along the lines of Donald and Newey (1999). It is "higher-order" in the sense that it is made up of the terms of the MSE that are of order N^{-2} and higher. It is "asymptotically optimal" in the sense that criterion function for use in practice approaches the true criterion function at a sufficient rate as sample size increases. It will be shown that the higher-order asymptotically optimal moment selection criterion function derived in this chapter performs very well in the selection of the optimal number of auxiliary moments for the GMM procedure, even in small sample sizes.

The GMM model is estimated using the Continuous Updating Estimator (henceforth CUE; Hansen, Heaton, and Yaron, 1996). The CUE is a special case of the class of generalized empirical likelihood (GEL) estimators (e.g. Qin and Lawless, 1994; Imbens, Spady, and Johnson, 1998; Newey and Smith, 2000). It has the nice properties of being a one-step procedure, being invariant to how the moment conditions are scaled, and having a closed-form solution, making MSE approximations easier. However, this benefit does not come without a cost. Hansen, Heaton, and Yaron (1996) show that in small samples the distribution of the CUE is relatively thick-tailed, even though it is first-order asymptotically efficient. Given the CUE closed-form solutions to the average treatment effect estimator and the Lagrange Multiplier estimator, the MSE of the average treatment effect estimator is approximated to the stochastic order N^{-2} (the second order) with asymptotic expansions of the two estimators. It is shown through simulations that in relevant small sample sizes (or larger) calculating the MSE to order N^{-2} yields a very accurate approximation to the true MSE. It is also shown that a higher-order asymptotically optimal approximation to the MSE formula performs well in terms of choosing the optimal estimator, even in small samples.

The format of the dissertation is as follows. Chapter 1 analyzes the case where the propensity score is constant and known, and where the covariate, X, is a scalar. The data generating process is that of an experimental setting where treatments and controls are assigned sequentially[1] (e.g. via a coin flip) and information on a scalar pre-treatment variable is obtained. Chapter 2 extends the analysis of Chapter 1 by allowing X to be a $D \times 1$ vector of pre-treatment variables. The ordering of the higher-order terms of X and its interactions may be an important issue in this case. Finally, Chapter 3 analyzes the case where the propensity score is unknown. This data generating process is that of a non-experimental setting. Research with this type of data is becoming more frequently observed in economics and, thus, examination of this case is valuable.

The format of this chapter is as follows. Section 1.2 sets up the model and discusses previous results and estimators. Section 1.3 presents the GMM framework and the estimator of this dissertation. Section 1.4 presents a simple example and evidence of an efficiency gain from this type of procedure. Section 1.5 presents the theoretical results using the asymptotic expansions. Section 1.6 details the sample approximation to the theoretical MSE formula and presents results of Monte Carlo simulations testing the procedure. Section 1.7 examines the properties of the procedure in a "more realistic"

[1] Hence, the observations are independent.

setting through simulations with various data generating processes calibrated to the experimental dataset of LaLonde (1986). It also applies the procedure to the actual dataset, as one would do in practice. Lastly, Section 1.8 concludes.

1.2 Setup of Treatment Effect Models

1.2.1 The Model

Suppose one has a random sample of data of size N from a population of interest. For each unit i in the sample, there is i.i.d. data $\{(Y_i, T_i, X_i)\}$ where Y_i is an outcome variable, T_i is a binary treatment indicator, and X_i is a scalar pre-treatment variable. X_i is assumed to be a scalar throughout the remainder of the chapter in order to focus attention on the properties of the method being developed and to abstract away from any ordering considerations with respect to power series and interactions[2]. Thus, the pre-treatment variable and polynomials thereof have a natural ordering as $\{X_i, X_i^2, \ldots, X_i^K\}$. The case where X_i is a vector will be considered in Chapter 2.

For each unit i, $T_i \in \{0,1\}$ where $T_i = 1$ means unit i received the treatment of interest and $T_i = 0$ means unit i was a control. The parameter of interest is the average treatment effect

$$\tau = E(Y(1) - Y(0)), \qquad (1)$$

[2] If X is a vector, all results still hold but the natural ordering is lost and one must compare the MSE of various subsets, interactions, and polynomials of X in a less ordered way.

where $Y(1)$ is the potential outcome when assigned to the treatment and $Y(0)$ is the potential outcome when assigned as a control. The fundamental problem is that of missing data. Each individual unit i has potential outcomes $(Y_i(1), Y_i(0))$. However, since each unit either receives the treatment or is a control, the econometrician only observes

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

1.2.2 Previous Results and Estimators

Suppose assignment to treatment is random as is the case in experimental studies. In this case

$$(Y_i(1), Y_i(0)) \perp T_i. \tag{2}$$

Given that the potential outcomes are independent of treatment status, the simple difference-in-averages estimator of τ,

$$\hat{\tau}_s = \bar{y}_1 - \bar{y}_0 = \frac{1}{\sum_{r=1}^{N} t_r}\left(\sum_{i=1}^{N} y_i \cdot t_i\right) - \frac{1}{\sum_{r=1}^{N}(1 - t_r)}\left(\sum_{i=1}^{N} y_i \cdot (1 - t_i)\right) \tag{3}$$

is unbiased[3]. In non-experimental studies, it is typically believed that assignment to treatment is not independent of treatment status. Thus, $\hat{\tau}_s$ is biased. Given data on pre-treatment variables upon which selection is based, the following identification assumption is made.

ASSUMPTION 1

$$(Y_i(1), Y_i(0)) \perp T_i \mid X_i$$

This is the well-known Unconfoundedness assumption or the Selection on Observables assumption of Rosenbaum and Rubin. Given this assumption, one can construct an unbiased estimator of τ because

$$\begin{aligned}
\tau &= E\big(Y(1) - Y(0)\big) \\
&= E_X\big(E\big(Y(1) - Y(0) \mid X\big)\big) \\
&= E_X\big(E\big(Y(1) \mid T = 1, X\big) - E\big(Y(0) \mid T = 0, X\big)\big) \\
&= E_X\big(E(Y \mid T = 1, X) - E(Y \mid T = 0, X)\big). \qquad . \qquad (4)
\end{aligned}$$

Both the first and second terms of the inner expectation are directly estimable with a variety of methods. The outer expectation then averages these results over the distribution of X. Hahn (1998) proposes two estimators of equation (4). For both estimators, the outer expectation becomes a sample average. For the first estimator, the inner expectation uses the observed value of Y_i and the non-parametrically imputed missing value $(1 - T_i)\cdot\hat{E}(\,Y_i \mid T_i = 1, X_i) + T_i \cdot \hat{E}\,(Y_i \mid T_i = 0, X_i)$. For the second estimator, the inner expectation uses both of the non-parametrically imputed values, $\hat{E}(\,Y_i \mid T_i = 1, X_i)$ and $\hat{E}(\,Y_i \mid T_i = 0, X_i)$. Hahn shows that both of these estimators are efficient in the sense that they reach the semi-parametric efficiency bound.

[3] HIR (2000) shows that $\hat{\tau}_s$ is still not efficient, given information on the pre-treatment variable X_i.

In practice, direct adjustment on X may not always be practical, depending on its dimensionality (assuming X is a vector for the moment). As a possible solution to this problem, Rosenbaum and Rubin (1983) show that Unconfoundedness implies

$$\left(Y_i(1), Y_i(0)\right) \perp T_i \mid p(X_i),$$

where $p(X_i) = \Pr(T_i = 1 \mid X_i)$ is the propensity score, the probability of treatment conditional on X_i. This result implies that an unbiased estimator of τ can be constructed with direct adjustment on the propensity score.

$$\begin{aligned}\tau &= E\left(Y(1) - Y(0)\right) \\ &= E_X\left(E\left(Y(1) - Y(0) \mid p(X)\right)\right) \\ &= E_X\left(E\left(Y(1) \mid T = 1, p(X)\right) - E\left(Y(0) \mid T = 0, p(X)\right)\right) \\ &= E_X\left(E\left(Y \mid T = 1, p(X)\right) - E\left(Y \mid T = 0, p(X)\right)\right). \qquad . \end{aligned} \tag{5}$$

The two terms of the inner expectation are still directly estimable, but they are conditioned on the scalar propensity score rather than the (possible) vector X. Hahn (1998) shows that knowledge of the true propensity score is ancillary for estimation of average treatment effects. In other words, the semi-parametric efficiency bound is the same whether the true propensity score is known or unknown. This is an asymptotic result. Chapter 3 will show through simulations that in finite samples the propensity score is not ancillary for estimation of average treatment effects.

Rosenbaum (1987) and Hahn suggest that an average treatment effect estimator that conditions on the true propensity score, as in equation (5), will not reach the semi-parametric efficiency bound. Hahn and HIR both provide examples where this holds true. The intuition is that conditioning on an estimate of the propensity score controls for pre-treatment variables, which affect the propensity score and the outcome, better than the true propensity score. This seemingly counterintuitive result is due to the fact that the estimated propensity score compensates for sample divergences from true probabilities[4]. As an example, suppose one has data from an experimental study of two hundred individuals. The probability of selection for treatment is .5 and there is information on one covariate, sex. The true population distribution of males and females is 50% for each and there are 100 males and 100 females in the study. Even though assignment to treatment is random, suppose that out of the 100 treated individuals, 90 end up being males. Thus, out of the 100 controls, 90 end up being females. Use of the true propensity score estimator, in this case $\hat{\tau}_s$, would tend to yield the average outcome on treated males minus the average outcome on control females. If there is a significant difference in the average treatment effect for males and females, this estimator will yield terribly misleading results. Adjusting for the non-parametric propensity score, .9 for males and .1 for females, will effectively compensate for these sample divergences from true distributional probabilities.

Rosenbaum (1987) and HIR discuss an alternative to direct adjustment on the propensity score. They show that an estimator that weights observations by the inverse of the true propensity score is unbiased for τ. Namely,

[4] See Rosenbaum (1987).

$$\hat{\tau}_{tp} = \frac{1}{N}\sum_{i=1}^{N}\left(\frac{y_i \cdot t_i}{p(x_i)} - \frac{y_i \cdot (1-t_i)}{1-p(x_i)}\right) \tag{6}$$

is unbiased where $p(X_i)$, the true propensity score, is known. Finally, HIR show that

$$\hat{\tau}_{HIR} = \frac{1}{N}\sum_{i=1}^{N}\left(\frac{y_i \cdot t_i}{\hat{p}(x_i)} - \frac{y_i \cdot (1-t_i)}{1-\hat{p}(x_i)}\right) \tag{7}$$

is consistent for τ and, unlike $\hat{\tau}_{tp}$, it reaches the semi-parametric efficiency bound asymptotically, where $\hat{p}(X_i)$ is a non-parametric estimate of the propensity score. Thus, both $\hat{\tau}_{tp}$ and $\hat{\tau}_{HIR}$ asymptotically remove all the bias associated with selection on the pre-treatment variable X, but $\hat{\tau}_{HIR}$ is more efficient. As seen in the example above, this efficiency result is true even in experimental studies when bias due to self-selection is not an issue.

1.3 GMM Framework

In this section, the general specification for the solution framework is laid out. The estimation problem is modeled within the Generalized Method of Moments when the true propensity score is known. Instead of weighting by the non-parametric propensity score, additional information is incorporated through auxiliary moment restrictions that capture the full effect of the non-parametric propensity score when the full set is employed. It is shown that the true propensity score estimator can be achieved by one unconditional moment restriction. Then, it is shown that an estimator with identical asymptotic properties to $\hat{\tau}_{HIR}$ can be achieved by adding one conditional moment restriction that reflects the econometrician's knowledge of the propensity score. The full inclusion of this conditional moment restriction achieves the same informative value as weighting by the non-parametric propensity score. Under suitable regularity conditions, this estimator can be achieved by including an infinite but countable number of unconditional moment restrictions. Next, the form of the small sample estimator, which uses a finite number of these auxiliary unconditional moment restrictions, is discussed. Finally, the closed-form solution of the average treatment effect estimator is solved.

1.3.1 Moment Restrictions and Relation to Previous Work

Given the setup above, the estimation procedure is framed in the context of a GMM system of moments. In particular, let

$$\psi(y,t,x,\tau)=\begin{pmatrix}\psi_1(y,t,x,\tau)\\ \psi_2(t,x)\end{pmatrix},$$

where

$$\psi_1(y,t,x,\tau)=\left(\frac{y\cdot t}{p(x)}-\frac{y\cdot(1-t)}{1-p(x)}\right)-\tau=V-\tau,$$

$$\psi_2(t,x)=\begin{pmatrix}t-p(x)\\ x(t-p(x))\\ \vdots\\ x^K(t-p(x))\end{pmatrix},$$

$$E(\psi(y,t,x,\tau))=0,$$

where $p(x)$ is the true propensity score. In this chapter, the propensity score is assumed to be constant $(p(X)=p)$ and known[5]. It will be seen below that as $K \to \infty$ at a suitable rate, $\psi_2(t,x)$ accounts for the conditional moment restriction $E(T-p(X)\mid X)=0$, reflecting one's knowledge of the propensity score. Let $\psi(y,t,x,\tau)=\psi$, $\psi_1(y,t,x,\tau)=\psi_1$, and $\psi_2(t,x)=\psi_2$ for simplicity. Also, let $M=K+2$ be the number of moments of ψ. Notice that this framework embeds the true propensity score estimator. If one lets $\psi=\psi_1$, there will be one moment restriction and one parameter to be estimated. Taking the sample average yields

$$\frac{1}{N}\sum_{i=1}^{N}\left(\frac{y_i \cdot t_i}{p(x_i)}-\frac{y_i \cdot (1-t_i)}{1-p(x_i)}-\tau\right)=0.$$

Solving this for τ yields the true propensity score estimator, $\hat{\tau}_{tp}$.

The motivation for the auxiliary conditional moment restriction comes from HIR, Chamberlain (1987), and Hellerstein and Imbens (1999). Hellerstein and Imbens show that efficiency can be improved by including auxiliary moment restrictions, ψ_2, that are correlated with the primary moment restrictions, ψ_1, even if ψ_2 does not include any unknown parameters[6]. Chamberlain shows that a conditional moment restriction of the form $E(Y-\beta X\mid X)=0$ implies $E(a(X)(Y-\beta X))=0$ for any function a. This

[5] $p(X)$ is written instead of p in the GMM model to reinforce that all of the theoretical results do not hinge on the propensity score being constant.

[6] It is only a function of T and X in our case.

unconditional moment restriction is satisfied for any a such that $E\left(a(X)^2\right)<\infty$ with an infinite but countable number of moment restrictions based on a power series of X,

$$E\begin{pmatrix} (Y-\beta X) \\ X(Y-\beta X) \\ X^2(Y-\beta X) \\ \vdots \end{pmatrix}=0.$$

This countable number of unconditional moment restrictions fully accounts for the information contained in the conditional moment restriction, provided some regularity and growth rate conditions hold. Donald, Imbens, and Newey (2001) discuss the necessary regularity conditions and provide the maximum and minimum growth rates for Empirical Likelihood (EL), GMM, and IV estimation while using splines, power series, or Fourier series of X.

HIR further motivate the conditional moment restriction in an example with a constant, known propensity score, $p(X)=p$, and a single binary pre-treatment variable, X. In their example, they show that inclusion of the two auxiliary moment restrictions implied by the two-point support of X results in obtaining the efficient HIR non-parametric propensity score estimator, $\hat{\tau}_{HIR}$.

PROPOSITION 1.1: *Suppose Assumptions 1 – 4 (2 – 4 listed below) hold. Let $p(X)$ be a general function of a discrete vector X. The estimator obtained by solving the full* *set[7] of auxiliary moment restrictions is the HIR non-parametric propensity score estimator, $\hat{\tau}_{HIR}$.*

Proof: See Appendix.

Thus, with discrete X, the estimator obtained from the GMM framework is asymptotically efficient and is, in fact, the HIR estimator. The above proof solves with the EL estimator rather than the CUE. However, since this example deals with asymptotic efficiency, this point is moot because both estimators have the same asymptotic distribution[8]. For continuous X, it is known that a general continuous distribution can be approximated arbitrarily closely by a discrete multinomial distribution[9]. Because of this, when X is continuous the estimator obtained as $K \rightarrow \infty$ in the GMM framework will have the same properties (consistency and asymptotic efficiency) as the HIR non-parametric propensity score estimator, $\hat{\tau}_{HIR}$, although the two need not be identical. This is due to two factors: the fact that a general continuous distribution can be approximated arbitrarily closely by a discrete multinomial distribution and the fact that as $K \rightarrow \infty$ at a suitable rate, the auxiliary *unconditional* moment restrictions fully account for the *conditional* moment restriction reflecting the auxiliary information about the conditional expectation of T.

[7] If X has J support points, then J auxiliary moment restrictions compose the full set.

[8] In this example, it was easier to solve with the EL estimator.

[9] See Chamberlain (1987).

It has been shown that the true propensity score weighted estimator is unbiased, but inefficient. It has also been shown that there can be an efficiency gain from taking account of additional information contained in the ψ_2 moments. The HIR estimator takes full account of this information and has been shown to reach the asymptotic efficiency bound. However, this estimator (and close approximations thereof) would not be optimal in small samples. K would be too large relative to sample size and inversions of $\psi_2\psi_2'$ would be difficult to impossible. In other words, the additional noise from the marginal auxiliary moment restrictions would be outweighing the gain of being closer to a perfect representation of the conditional moment restriction, $E(T-p(X)\mid X)=0$. This suggests that in small samples there can be an efficiency gain to moving away from the true propensity score estimator and including additional moment restrictions <u>until</u> the efficiency loss effect begins to dominate the efficiency gain as what is similar to the HIR estimator is approached.

1.3.2 Estimation Procedure with the CUE

ASSUMPTION 2

The support of X is a compact interval on $\mathbb{R}$.

ASSUMPTION 3

$E\left(Y(1)^2\right)<\infty$ *and* $E\left(Y(0)^2\right)<\infty$.

ASSUMPTION 4

$0<\underline{p}<p(X)<\overline{p}<1 \text{ where } p(X)=\Pr(T=1\mid X=x)$.

Assumption 2 is necessary for the regularity conditions of the power series of X to hold and for the derivation of the asymptotic expansions of the average treatment effect estimator. Assumptions 3 and 4 are required for the derivation of the MSE properties of the average treatment effect estimator.

Given the setup above, the GMM moment conditions are solved using the CUE. As mentioned previously, the CUE is a special case of the class of generalized empirical likelihood (GEL) estimators (e.g. Qin and Lawless, 1994; Imbens, Spady, and Johnson, 1998; Newey and Smith, 2000). It is obtained by simultaneously minimizing the GMM criterion function (including the weight matrix) with respect to τ. In other words

$$\hat{\tau}_{CUE}=\operatorname*{argmin}_{\tau\in T}\left(\frac{1}{N}\sum_{r=1}^{N}\psi_r(\tau)\right)'\left(\frac{1}{N}\sum_{s=1}^{N}\psi_s(\tau)\psi_s(\tau)'\right)^{-1}\left(\frac{1}{N}\sum_{i=1}^{N}\psi_i(\tau)\right).$$

Newey and Smith provide the first-order conditions for GEL estimators. With the moment restrictions of this paper, their first-order conditions turn into[10]

$$\begin{pmatrix}\sum_i \psi_{1i}\left(1-\lambda'\psi_{2i}\right)\\ \sum_i \psi_{2i}\left(1-\lambda'\psi_{2i}\right)\end{pmatrix}=0, \tag{8}$$

[10] The CUE is a special case in which Newey and Smith's $\rho(v)$ function is a quadratic. Specifically, let $\rho(v)=2v-v^2,\ 3v-\frac{3}{2}v^2,\ 4v-\frac{4}{2}v^2$, or...

where λ is a $(M-1\times1)$ vector of Lagrange Multipliers[11]. As mentioned earlier, the CUE procedure has the three nice properties of being a one-step procedure, being invariant to the scaling of the moment conditions, and having a closed-form solution. Using a one-step procedure is nice because it removes any arbitrariness associated with choosing a consistently estimated weight matrix. The closed-form solution will be extremely helpful in Section 1.5 when asymptotic expansions of $\hat{\lambda}$ and $\hat{\tau}$ are derived. To solve using the CUE, one simultaneously solves the M equations of (8). This yields

$$\hat{\lambda}=\left(\frac{1}{N}\sum_{r}\psi_{2r}\psi_{2r}'\right)^{-1}\frac{1}{N}\sum_{i}\psi_{2i}, \tag{9}$$

$$\hat{\tau}=\frac{\frac{1}{N}\sum_{i}Y_i\left(1-\hat{\lambda}'\psi_{2i}\right)}{\frac{1}{N}\sum_{r}\left(1-\hat{\lambda}'\psi_{2r}\right)}. \tag{10}$$

Equations (9) and (10) contain the closed-form solutions for $\hat{\lambda}$ and $\hat{\tau}$. In Section 1.5 approximate MSE criteria for $\hat{\tau}$ will be derived using the leading terms in asymptotic expansions of these two estimators.

1.4 A Simple Example

Let

$$Y_i=.5+2T_i+\beta_2X_i+\varepsilon_i,$$

$$\varepsilon_i\sim N(0,1),$$

[11] λ is of dimension $M-1$ because the Lagrange Multiplier for the ψ_1 moment restriction is equal to zero so it drops out of all equations and the estimation procedure.

$$X_i \sim UNIF(-1,1),$$

$$p(X_i) = \frac{1}{2}.$$

In this data generating process, assignment to treatment is random and the treatment effect is constant for all individuals. Therefore, the true average treatment effect equals the individual treatment effect, $\tau = 2$. HIR have shown that $\hat{\tau}_{tp}$ is asymptotically inefficient. It has also been suggested that the HIR non-parametric propensity score estimator is not the optimal estimator in small samples. In Figure 1.1, $\beta_2 = 1$ and N = 25. The true MSE of $\hat{\tau}$ is calculated by taking the sample MSE of 10,000 estimates of τ and plotting these MSEs over a range of K values. It can be see that the true propensity score estimator, $\hat{\tau}_{tp}$ (henceforth called the $K = -1$ estimator), performs particularly poorly with a MSE of .594. It is straightforward to show that when the propensity score is constant, the GMM estimator that sets $K = 0$ corresponds to the simple difference-in-averages estimator, $\hat{\tau}_s$. This estimator is the most commonly used unbiased estimator in experimental studies. It has a MSE of .220. It contains a 63% MSE reduction relative to the true propensity score estimator of $K = -1$. The intuition behind this result is relatively simple. The difference-in-averages estimator corrects for sample divergences of the empirical selection probability from the true selection probability due to the (assumed) sequential selection design of the experiment. If the experiment has a non-sequential selection design, the true propensity score estimator will equal the difference-in-averages estimator. It is worth noting that the MSE reduction of $K = 0$ relative to

$K=-1$ was in the 60% – 80% range for all data generating processes examined[12]. The estimator with $K=1$ has a MSE of .177. It contains a 20% MSE reduction relative to the difference-in-averages estimator ($K=0$). This MSE reduction is due to the fact that the $x(t-p)$ moment corrects for sample divergences of the empirical distribution of X from its true population distribution for the treated and for the controls. The relative importance of this correction depends upon the magnitude of the effect of X in the outcome equation (in this example the effect is β_2)[13]. Finally, the estimator with $K=6$ has a MSE of .233. This estimator can be thought of as being similar to the HIR non-parametric propensity score estimator due to the fact that it incorporates a large amount of knowledge of the propensity score in the form of auxiliary moment restrictions. The $K=1$ estimator contains a 24% MSE reduction relative to the $K=6$ estimator. This is evidence that, due to the small sample size of N = 25, the noise created by the addition of moments past $K=1$ is significant, and the relative benefit of these moments is small.

Figure 1.2 shows the MSE graph from the same data generating process, except N = 150. In this case, again there exists a 63% MSE reduction in going from the $K=-1$ estimator to the $K=0$ estimator. There exists a 24% MSE reduction in going from the $K=0$ estimator to the $K=1$ estimator (similar to when N = 25). Lastly, there exists only a 2% MSE reduction in going from the $K=6$ estimator to the $K=1$ estimator.

[12] Even when $\beta_2=0, Var(\varepsilon)=0$, and N = large.

[13] Under a general data generating process or a general $p(X)$, this moment will partially correct for sample divergences. Higher-order moments (of X) may be optimal if higher-order terms of X are strong in the outcome equation or enter the propensity score.

Thus, $\hat{\tau}_{tp}$ and $\hat{\tau}_s$ remain significantly sub-optimal and the efficiency loss from using a large K decreases. This is evidence that $\hat{\tau}_{HIR}$[14] will be optimal in large samples since the marginal cost of noise from marginal moments decreases with sample size.

Figures 1.3 and 1.4 show MSE plots similar to Figures 1.1 and 1.2 except β_2 has been reduced to $\beta_2 = .3$. Thus, the effect of X in the outcome equation has been reduced. Therefore, all else equal, the benefit from auxiliary moments containing powers of X should be reduced as well. In Figure 1.3, N = 25 and the optimal estimator is the $K = 0$ estimator. The $K = 0$ estimator contains a 67% MSE reduction relative to the $K = -1$ estimator. The MSE of the $K = 0$ estimator is .173, while the MSE of the $K = 1$ estimator is .179. This represents only a 3% MSE reduction in the $K = 0$ estimator relative to the $K = 1$ estimator. This is due to the fact that $\beta_2 = .3$ is small enough in magnitude such that the optimal K is $K = 0$. However, β_2 is close to the critical value where the optimal K becomes $K = 1$. The MSE reduction of the optimal $K = 0$ estimator relative to the $K = 6$ estimator is 25%. This is slightly larger than the MSE reduction of Figure 1.1 due to the reduction in the benefit of auxiliary moments containing powers of X.

Figure 1.4 plots the MSE graph for $\beta_2 = .3$ and N = 150. Now, the $K = 1$ estimator becomes optimal. The increase in sample size was large enough to mitigate the effect of the additional noise of the $K = 1$ moment. Thus, the optimal estimator switched from the $K = 0$ to the $K = 1$ estimator. Again, there is a large MSE reduction in going

[14] Or an estimator with large K.

from the $K = -1$ estimator to the $K = 0$ estimator, 68%. There is a 4% MSE reduction in going from the $K = 0$ estimator to the $K = 1$ estimator. Finally, there is a 5% MSE reduction in going from the $K = 6$ estimator to the $K = 1$ estimator. 5% is relatively small, due to the large sample size. However, it is larger than the MSE reduction of Figure 1.2 (2%) again due to the reduction in the power of X in the outcome equation.

When the propensity score is constant and the outcome equation is linear in X with no higher-order polynomials, the maximum optimal K is $K = 1$. This is due to the fact that if the linear term X is the highest-order polynomial entering the outcome equation, there can be no further gain once X is accounted for in the auxiliary moments. Figures 1.5 and 1.6 show results where the optimal K can be greater than one. In this case, the propensity score is a logit probability, including a linear and quadratic term of X. Figure 1.5 shows that the optimal estimator is the $K = 1$ estimator when $\beta_2 = 1$. Figure 1.6 shows that when β_2 is increased to $\beta_2 = 8$, the $K = 3$ estimator becomes optimal.

The simple example of this section brings to light a few important points. One, the true propensity score estimator performs extremely poorly for all data generating processes. It has a MSE that is at least 250% greater than the optimal estimator. Two, when X is strong enough in the outcome equation, a non-trivial MSE reduction[15] can obtain by incorporating additional auxiliary moments ($K = 1$ in this case) relative to the difference-in-averages estimator ($K = 0$). Three, when sample size is small enough, the

[15] 20% - 24% in our examples.

optimal estimator can contain a non-trivial MSE reduction[16] relative to the large K estimator, which is similar to the HIR non-parametric propensity score estimator.

1.5 Asymptotic Expansions

1.5.1 Higher-order MSE and $S(K)$ formulae

In this section, asymptotic expansions for the two estimators, $\hat{\lambda}$ and $\hat{\tau}$ are derived. This is done by using methods similar to Newey and Smith (2000) and Donald and Newey (1999). First, the asymptotic expansion for $\hat{\lambda}$ is derived. Then, the leading terms in the asymptotic expansion of $\hat{\lambda}$ are used to derive the asymptotic expansion of $\hat{\tau}-\tau$. $\hat{\tau}-\tau$ can be written as the sum of the leading terms of its asymptotic expansion[17]. Thus, the approximate MSE of $\hat{\tau}$ is the expectation of the square of these leading terms.

Relevant Notation

i. $\psi_{2ij} = j^{th}$ element of ψ_{2i}

ii. $\Omega = E\left(\psi_{2i}\psi_{2i}'\right)$

iii. $\omega_{ij} = (i, j)$ element of Ω

iv. $\omega_{ij}^{-1} = (i, j)$ element of Ω^{-1}

v. $\mathrm{H} = \frac{1}{n}\sum_{i}\left(\psi_{2i}\psi_{2i}' - \Omega\right)$

[16] 24% - 25% in our examples.

vi. $\sigma_{V\Psi_2} = E\left((V_i - \tau)\psi_{2i}\right)$

vii. $\sigma_{V\Psi_{2j}} = j^{th}$ element of $\sigma_{V\Psi_2}$

viii. $\sigma_V^{\,2} = E\left((V_i - \tau)^2\right)$

With the solution for $\hat{\tau}$ in equation (10), the goal is to derive its MSE for various K. This will allow the optimal K to be chosen. However, given the functional form of $\hat{\tau}$, derivation of the MSE analytically is not possible. Thus, asymptotic expansions of the two estimators, $\hat{\lambda}$ and $\hat{\tau}$, become useful. There are three relevant expansions of the components of $\hat{\lambda}$.

ASYMPTOTIC EXPANSION 1.1: *Suppose Assumptions 1 –4 hold.*

(i) $$\frac{1}{N}\sum_i \psi_{2i} = O_p\left(N^{-1/2}\right), \qquad (11)$$

(ii) $$\mathrm{H} = \frac{1}{N}\sum_i \left(\psi_{2i}\psi_{2i}{}' - \Omega\right) = O_p\left(N^{-1/2}\right), \qquad (12)$$

(iii) $$\left(\frac{1}{N}\sum_i \psi_{2i}\psi_{2i}{}'\right)^{-1} = \Omega^{-1} - \Omega^{-1}\mathrm{H}\Omega^{-1} + \Omega^{-1}\mathrm{H}\Omega^{-1}\mathrm{H}\Omega^{-1} + o_p\left(N^{-1}\right), \qquad (13)$$

Proof: See Appendix.

LEMMA 1.1: *Suppose Assumptions 1 –4 hold.*

[17] The terms of high order of convergence (or low rate of convergence).

$$\hat{\lambda} = T_{N\lambda} + R_{N\lambda} + S_{N\lambda} + o_p\left(N^{-3/2}\right), \tag{14}$$

where $T_{N\lambda} = \Omega^{-1}\frac{1}{N}\sum_i \psi_{2i} = O_p\left(N^{-1/2}\right)$, $R_{N\lambda} = -\Omega^{-1}\mathrm{H}\Omega^{-1}\left(\frac{1}{N}\sum_i \psi_{2i}\right) = O_p\left(N^{-1}\right)$, *and*

$S_{N\lambda} = \Omega^{-1}\mathrm{H}\Omega^{-1}\mathrm{H}\Omega^{-1}\left(\frac{1}{N}\sum_i \psi_{2i}\right) = O_p\left(N^{-3/2}\right)$.

Proof: See Appendix.

Next, asymptotic expansions for the three factors of $\hat{\tau} - \tau$ are derived. This is done using the result of equation (14).

$$\hat{\tau} - \tau = \frac{\frac{1}{N}\sum_i V_i\left(1 - \hat{\lambda}'\psi_{2i}\right)}{\frac{1}{N}\sum_r\left(1 - \hat{\lambda}'\psi_{2r}\right)} - \tau$$

$$= \left(\frac{1}{N}\sum_r (V_r - \tau) - \frac{1}{N}\sum_s (V_s - \tau)\hat{\lambda}'\psi_{2s}\right)\left(\frac{1}{N}\sum_i\left(1 - \hat{\lambda}'\psi_{2i}\right)\right)^{-1}. \tag{15}$$

ASYMPTOTIC EXPANSION 1.2: *Suppose Assumptions 1 – 4 hold.*

(i) $$\frac{1}{N}\sum_i (V_i - \tau) = O_p\left(N^{-1/2}\right), \tag{16}$$

(ii) $$\frac{1}{N}\sum_i (V_i - \tau)\psi_{2i} = \sigma_{V\psi_2} + O_p\left(N^{-1/2}\right), \tag{17}$$

(iii) $$\frac{1}{N}\sum_i (V_i - \tau)\hat{\lambda}'\psi_{2i} = \hat{\lambda}'\sigma_{V\psi_2} + R_{N\tau} + S_{N\tau} + o_p\left(N^{-3/2}\right), \tag{18}$$

(iv) $$\left(\frac{1}{N}\sum_i\left(1-\hat{\lambda}'\psi_{2i}\right)\right)^{-1} = 1+\hat{\lambda}'\left(\frac{1}{N}\sum_i\psi_{2i}\right)+o_p\left(N^{-3/2}\right), \tag{19}$$

where $R_{N\tau} = \left(\Omega^{-1}\frac{1}{N}\sum_s\psi_{2s}\right)'\left(\frac{1}{N}\sum_i\left((V_i-\tau)\psi_{2i}-\sigma_{V\psi_2}\right)\right) = O_p\left(N^{-1}\right)$ *and*

$$S_{N\tau} = -\left(\Omega^{-1}\mathrm{H}\Omega^{-1}\left(\frac{1}{N}\sum_s\psi_{2s}\right)\right)'\left(\frac{1}{N}\sum_i\left((V_i-\tau)\psi_{2i}-\sigma_{V\psi_2}\right)\right) = O_p\left(N^{-3/2}\right).$$

Proof: See Appendix.

To calculate the MSE of $\hat{\tau}$, the closed-form solution for $\hat{\tau}-\tau$ is decomposed into the sum of terms of stochastic order $N^{-1/2}$, N^{-1}, and $N^{-3/2}$. This is done by simplifying the closed-form solution for $\hat{\tau}-\tau$ using the asymptotic expansion for $\hat{\lambda}$ [equation (14)] and the asymptotic expansions for $\hat{\tau}-\tau$ [equations (16) – (19)].

THEOREM 1.1: *Suppose Assumptions 1 - 4 hold.*

$$\hat{\tau}-\tau = T_{N1} + T_{N2} + R_{N1} + R_{N2} + S_{N1} + S_{N2} + S_{N3} + S_{N4} + o_p\left(N^{-3/2}\right), \tag{20}$$

where

$$T_{N1} = \frac{1}{N}\sum_i(V_i-\tau) = O_p\left(N^{-1/2}\right),$$

$$T_{N2} = -\sigma_{V\psi_2}{}'\Omega^{-1}\left(\frac{1}{N}\sum_i\psi_{2i}\right) = O_p\left(N^{-1/2}\right),$$

$$R_{N1} = \left(\frac{1}{N}\sum_i\psi_{2i}\right)'\Omega^{-1}\mathrm{H}\Omega^{-1}\sigma_{V\psi_2} = O_p\left(N^{-1}\right),$$

$$R_{N2} = -\left(\frac{1}{N}\sum_{s}\psi_{2s}\right)'\Omega^{-1}\left(\frac{1}{N}\sum_{i}\left((V_i-\tau)\psi_{2i}-\sigma_{V\psi_2}\right)\right) = O_p\left(N^{-1}\right).$$

$$S_{N1} = -\left(\frac{1}{N}\sum_{i}\psi_{2i}\right)'\Omega^{-1}\mathrm{H}\Omega^{-1}\mathrm{H}\Omega^{-1}\sigma_{V\psi_2} = O_p\left(N^{-3/2}\right)$$

$$S_{N2} = \left(\frac{1}{N}\sum_{s}\psi_{2s}\right)'\Omega^{-1}H\Omega^{-1}\left(\frac{1}{N}\sum_{i}\left((V_i-\tau)\psi_{2i}-\sigma_{V\psi_2}\right)\right) = O_p\left(N^{-3/2}\right)$$

$$S_{N3} = \left(\frac{1}{N}\sum_{r}(V_r-\tau)\right)\left(\frac{1}{N}\sum_{s}\psi_{2s}\right)'\Omega^{-1}\left(\frac{1}{N}\sum_{i}\psi_{2i}\right) = O_p\left(N^{-3/2}\right)$$

$$S_{N4} = -\sigma_{V\psi_2}{}'\Omega^{-1}\left(\frac{1}{N}\sum_{r}\psi_{2r}\right)\left(\frac{1}{N}\sum_{s}\psi_{2s}\right)'\Omega^{-1}\left(\frac{1}{N}\sum_{i}\psi_{2i}\right) = O_p\left(N^{-3/2}\right)$$

Proof: See Appendix.

Typically, when approximating a statistic using asymptotic expansions econometricians use all terms of the two highest orders ($N^{-1/2}$ and N^{-1} in our case). For MSE approximations, they typically keep the terms of the square of the highest order and higher ($N^{-1}N^{-1} = N^{-2}$ or above in our case). Therefore, using (20), the approximate MSE of $\hat{\tau}$ is calculated using the components of this MSE that are of stochastic order N^{-2} and above. Hence, the interactions of $R_{N\cdot}$ with $S_{N\cdot}$ and the interactions of $S_{N\cdot}$ with itself are ignored for these are of stochastic order lower than N^{-2}.

COROLLARY 1.1: *The higher-order MSE of $\hat{\tau}$ is:*

$$
\begin{aligned}
MSE(\hat{\tau}) = E\left((\hat{\tau}-\tau)^2\right) = E\left((T_{N1} + T_{N2} + R_{N1} + R_{N2} + S_{N1} + S_{N2} + S_{N3} + S_{N4})^2\right) \qquad (21) \\
= E\left(T_{N1}^{\ 2}\right) + E\left(T_{N2}^{\ 2}\right) + E\left(R_{N1}^{\ 2}\right) + E\left(R_{N2}^{\ 2}\right) + 2E(T_{N1}T_{N2}) \\
+2E(T_{N1}R_{N1}) + 2E(T_{N2}R_{N1}) + 2E(T_{N1}R_{N2}) + 2E(T_{N2}R_{N2}) + 2E(R_{N1}R_{N2}) \\
+2E(T_{N1}S_{N1}) + 2E(T_{N2}S_{N1}) + 2E(T_{N1}S_{N2}) + 2E(T_{N2}S_{N2}) \\
+2E(T_{N1}S_{N3}) + 2E(T_{N2}S_{N3}) + 2E(T_{N1}S_{N4}) + 2E(T_{N2}S_{N4}) .
\end{aligned}
$$

Proof: See Appendix. Note: All of the expectations above are derived in the appendix.

Notice that all of the components making up the MSE of $\hat{\tau}$ in equation (21) are of stochastic order N^{-2} or higher. Also notice (in the appendix) that the $E\left(T_{N1}^{\ 2}\right)$ term does not depend on K and can therefore be excluded from the criterion function without affecting the optimal choice of K. This leads to the following definition of the criterion function which, when minimized, corresponds to minimizing the MSE.

DEFINITION: *The higher-order MSE criterion function of $\hat{\tau}$ is:*

$$
\begin{aligned}
S(K) = E\left(T_{N2}^{\ 2}\right) + E\left(R_{N1}^{\ 2}\right) + E\left(R_{N2}^{\ 2}\right) + 2E(T_{N1}T_{N2}) \quad . \qquad (22) \\
+2E(T_{N1}R_{N1}) + 2E(T_{N2}R_{N1}) + 2E(T_{N1}R_{N2}) + 2E(T_{N2}R_{N2}) + 2E(R_{N1}R_{N2}) \\
+2E(T_{N1}S_{N1}) + 2E(T_{N2}S_{N1}) + 2E(T_{N1}S_{N2}) + 2E(T_{N2}S_{N2}) \\
+2E(T_{N1}S_{N3}) + 2E(T_{N2}S_{N3}) + 2E(T_{N1}S_{N4}) + 2E(T_{N2}S_{N4}) ,
\end{aligned}
$$

Figures 1.7 and 1.8 show true MSE plots along with MSE plots derived from the higher-order formula above. The true MSE is derived in the same way as in Section 1.4. The higher-order formula MSE is derived from true variances and covariances, which are estimated as sample variances and covariances from a very large sample size of 100,000. From Figure 1.7, it is seen that the higher-order formula provides a tight fit to the truth, even with a small sample size, N = 25. The importance of the second-order terms can also be seen in this figure. The dashed and dotted line is the MSE obtained from only the first-order terms of the formula. The first-order formula is always decreasing as K increases. It provides an accurate approximation only up to $K = 0$ or $K = 1$, then diverges significantly after. From Figure 1.8, one can see that the formula works extremely well in larger sample sizes, N = 150 in this case. Also, notice that the first-order formula provides a relatively tight fit here as well. This is expected and is due to the fact that with larger sample sizes the lower-order terms are very close to zero. Thus, the lower-order terms are non-trivial with respect to the performance of the higher-order MSE formula and, when included, the approximated MSE formula yields a very tight fit to the true MSE. Plots containing the true $S(K)$ and the higher-order formula $S(K)$ are identical to the MSE plots except a constant is subtracted from the MSE plots such that $S(-1) = 0$.

1.5.2 Higher-order Asymptotic Optimality

In this section, the properties of the MSE criterion function, $S(K)$ are considered. The rule for selecting K is defined and it is shown that this rule is higher-order asymptotically optimal. The analysis follows that of Donald and Newey (1999) and Li (1987).

DEFINITION: *Let $\hat{S}(K)$ be the sample analog of $S(K)$. It is composed of sample quantities $\hat{E}(\cdot)$, which equal the sample analog of $E(\cdot)$.*

The goal is to choose K such that the MSE of $\hat{\tau}$ is minimized. This corresponds to choosing K such that $S(K)$ is minimized. Thus, for use in practice, the optimal choice for the auxiliary moment restrictions is defined to be

$$\hat{K} = \operatorname*{argmin}_{K} \hat{S}(K).$$

When X is a scalar, this minimum is relative to an index set of K values. In Chapter 2 when X is a vector, the definition of the set of possible K values may become an important consideration. Donald and Newey (1999) make the following definition.

DEFINITION: *A method of selecting K is defined to be "higher-order asymptotically optimal with respect to the criterion $S(K)$" if it can be shown that:*

$$\frac{S(\hat{K})}{\inf_{K} S(K)} \xrightarrow{p} 1$$

The term "higher-order" derives from the fact that the MSE formula is a higher-order approximation to the true MSE, and hence the $S(K)$ formula is a higher-order approximation to the true $S(K)$.

LEMMA 1.2: *Suppose Assumptions 1 – 4 hold.*

$$\hat{\sigma}_V^2 - \sigma_V^2 = O_p\left(N^{-1/2}\right),$$

$$\hat{\sigma}_{V\psi_2} - \sigma_{V\psi_2} = O_p\left(N^{-1/2}\right),$$

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

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

$$\widehat{Cov}\left((V_i - \tau)\psi_{2il}, \psi_{2it}\psi_{2ij}\right) - Cov\left((V_i - \tau)\psi_{2il}, \psi_{2it}\psi_{2ij}\right) = O_p\left(N^{-1/2}\right),$$

$$\widehat{Cov}\left(\psi_{2ip}\psi_{2il}, \psi_{2it}\psi_{2ij}\right) - Cov\left(\psi_{2ip}\psi_{2il}, \psi_{2it}\psi_{2ij}\right) = O_p\left(N^{-1/2}\right),$$

$$\widehat{Cov}\left((V_i - \tau)\psi_{2it}, (V_i - \tau)\psi_{2ij}\right) - Cov\left((V_i - \tau)\psi_{2it}, (V_i - \tau)\psi_{2ij}\right) = O_p\left(N^{-1/2}\right),$$

$$\widehat{Cov}\left(\psi_{2il}, \psi_{2it}\psi_{2ij}\right) - Cov\left(\psi_{2il}, \psi_{2it}\psi_{2ij}\right) = O_p\left(N^{-1/2}\right),$$

$$\widehat{Cov}\left(\psi_{2it}, (V_i - \tau)\psi_{2ij}\right) - Cov\left(\psi_{2it}, (V_i - \tau)\psi_{2ij}\right) = O_p\left(N^{-1/2}\right),$$

$$\widehat{Cov}\left((V_i - \tau), \psi_{2it}\psi_{2ij}\right) - Cov\left((V_i - \tau), \psi_{2it}\psi_{2ij}\right) = O_p\left(N^{-1/2}\right),$$

$$\widehat{Cov}\left(V_i-\tau,\left(V_i-\tau\right)\psi_{2ij}\right)-Cov\left(V_i-\tau,\left(V_i-\tau\right)\psi_{2ij}\right)=O_p\left(N^{-1/2}\right),$$

Proof: See Appendix.

The estimated variances and covariances above are sample analogs to the true variances and covariances. Given Lemma 1.2, the higher-order asymptotic optimality of the criterion function $\hat{S}(K)$ can be proven.

THEOREM 1.2: *Suppose Assumption 5 holds. The rule "select the value of K that minimizes $\hat{S}(K)$ " is higher-order asymptotically optimal with respect to the true minimization criterion $S(K)$.*

Proof: See Appendix.

1.6 Monte Carlo Simulations with One Covariate

The previous section presented simulation evidence of the goodness-of-fit of the higher-order MSE approximation formula and, subsequently, the higher-order criterion function, $S(K)$. It also showed that the criterion function for use in practice, $\hat{S}(K)$, is higher-order asymptotically optimal with respect to $S(K)$. The following two sections provide evidence of the performance the GMM procedure and $\hat{S}(K)$ with actual finite sample datasets. This section examines the performance of the procedure relative to data generating processes similar to those of Sections 1.4 and 1.5. The procedure is executed for 1,000 different Monte Carlo samples. The distribution of the optimal K, $\hat{K}$, over these 1,000 samples is plotted and analyzed. Also, the mean of $\hat{S}(K)$ over the 1,000 samples is plotted and compared to the true $S(K)$. The two are not convergent for all K values[18], however their minima tend to be the same K value. This is the most important property with respect to optimal selection of K.

Figures 1.9 and 1.10 plot results for the data generating process where $y = .5 + 2t + x + .4tx + \varepsilon$ and N = 25. All other parameters are the same as they were previously. Note that an interaction term has been added, however, since X is already "strong" in the outcome equation, it doesn't affect the optimal K. Inclusion of this interaction term affects the graphs in a trivial way. Figure 1.10 shows that the optimal K value is $K = 1$. Also, relative to $K = -1$, $K = 0$, and $K = 6$, the $K = 1$ estimator contains a significant reduction in the value of $S(K)$. Accordingly, the PDF of the optimal K distribution has a relatively tight fit around its mode at $K = 1$ with a value of close to .4. The right tail of the distribution tends to be thicker due to the fact that the true $S(K)$ is closer to the optimal $S(K)$ for K values greater than $K = 1$ versus those less than $K = 1$. Also, $K = -1$ is never chosen. This is a nice result given the estimator's high MSE. Figure 1.10 shows that the mean of $\hat{S}(K)$ is similar to the true $S(K)$ and their $\underset{K}{\operatorname{argmin}}$ is the same.

[18] As sample size increases, the two tend to converge.

Figures 1.11 and 1.12 plot results from the same data generating process except N = 150. In this case, the $K = 1$ estimator is still the true optimal estimator. Also, there still exists a significant $S(K)$ reduction for the $K = 1$ estimator relative to the $K = -1$ and $K = 0$ estimators. However, the significant difference between this data generating process and the previous one is that the $S(K)$ reduction for the $K = 1$ estimator relative to the $K = 6$ estimator is small, approximately 2%. Therefore, one would expect the PDF of the distribution of optimal K values to have its mode at $K = 1$ and be extremely thick-tailed to the right. Figure 1.11 shows this to be exactly the case, except technically the mode of the distribution is at $K = 6$. There exists a local maximum at $K = 1$ with the global maximum at $K = 6$. This is due to the fact that the MSEs are so close together and the sample size is large, so there is little difference between the $K = 1$ through $K = 6$ estimators. It is also readily observed that the $K = -1$ and $K = 0$ estimators are never chosen. This is not surprising and rather intuitive given the significant divergence in MSE of the two estimators from all other estimators with higher K values. It is interesting to note that the commonly used difference-in-averages estimator would be a very poor estimator to choose in this case, and the procedure never chooses it! Finally, as expected, Figure 1.12 shows us that the mean of $\hat{S}(K)$ tends to converge to the true $S(K)$ as sample size gets large.

The results of this section show that the procedure performs relatively well for the data generating processes considered. All of the results conformed to basic underlying intuition. Next, "more realistic" data generating processes will be considered. They are based off of the empirical distribution of a subset LaLonde's experimental dataset (1987).

1.7 Application to Experimental Data of LaLonde (1986)

In this section, the GMM procedure is applied to various calibrated data generating processes of the LaLonde (1986) experimental dataset and then to the actual dataset itself. The data generating process is calibrated to the empirical distribution of the relevant variables in this dataset. The results are fundamentally and intuitively similar to those of Section 1.6, however they should be more credible given that the data generating process is more similar to a real life data generation mechanism.

The dataset used in LaLonde's paper was gathered in the 1970's by the National Supported Work Demonstration (NSW). This was a temporary employment program designed to help disadvantaged workers by improving their job skills. In this experimental study, applicants were randomly assigned to either treatment or control status. Pre-treatment data was collected and the applicant's earnings were monitored over the next few years. Table 1.1 shows the sample means and standard deviations for the outcome variable (Earnings in 1978), the treatment variable, and all pre-treatment variables of a random subset of the NSW AFDC participants of LaLonde's dataset. Means and standard deviations are also shown separately for the treated and for the controls. Since selection into treatment status was truly random, the sample means and standard deviations of the pre-treatment variables (Earnings in 1975 and below) should be similar and non-statistically different from each other. This is observed to be the case.

Table 1.2 shows the OLS results when Earnings in 1978 is regressed on the treatment variable and all pre-treatment variables. The results are shown for non-standardized regressors and standardized regressors. It has been shown that the magnitude of a pre-treatment variable in the outcome equation is important with respect to moment selection. The standardized regression results are useful in gaining intuition on this magnitude since all regressors have identical, unit variance. One can see that Earnings in 1975 is the fourth "strongest" pre-treatment variable according to this specification. It was chosen as the scalar covariate in this chapter because in any earnings study it seems sensible to control for current earnings.

A Monte Carlo dataset is calibrated to the LaLonde dataset according to the following rules. Let the observed scalar covariate x_L be Earnings in 1975. Let (y_L, t_L) be the observed outcome vector and treatment vector from LaLonde's dataset, respectively. Let the propensity score equal the empirical mean of the treatment vector t_L, $p = \frac{1}{N}\sum_{i=1}^{N} t_{Li} = .416$. Run the OLS regression of

$$y_L = \beta_0 + \beta_1 t_L + \beta_2 x_L + \beta_3 t_L x_L + \varepsilon .$$

Let $\hat{\beta}$ be the resulting vector of parameter estimates and $\hat{\sigma}^2 = \hat{\varepsilon}'\hat{\varepsilon} / (N-4)$. Draw x from the empirical distribution of x_L. Draw t from a $BIN(1,p)$ distribution. Draw v from a $N(0,\hat{\sigma}^2)$ distribution. Finally, let

$$y = \hat{\beta}_0 + \hat{\beta}_1 t + \hat{\beta}_2 x + \hat{\beta}_3 tx + v .$$

In what follows, this calibration routine is used to examine the properties of the $\hat{S}(K)$ criterion. The simulated sample size is adjusted from N = 25 to N = 150 to N = 445 (LaLonde's sample size), and the $(\hat{\beta}_2, \hat{\beta}_3)$ values are adjusted to see how they affect the graphs.

Figures 1.13 and 1.14 show a PDF graph and $S(K)$ plot for a simulation sample size of N = 25. Notice that the true $S(K)$ reaches a minimum at $K = 0$ and the PDF of the optimal K values obtains its mode at $K = 0$ with a frequency of approximately .65. Until now, the choice set of K has been allowed to extend to a maximum of $K = 6$. However, with the current calibrated data generating process the maximum K value had to be restricted to something smaller. This is solely because the Earnings in 1975 variable is 65% zeros. In practice, if one of the 1,000 fake datasets has an x with a large amount of zeros, then x will be close to a linear combination of its powers and the routine will fail or return NaNs. The more powers of x allowed, the more common this occurrence will be. In contrast, the larger the simulated sample size, the less common this occurrence will be. Therefore, the maximum K value is restricted to $K = 2$ when N = 25, $K = 4$ when N = 150, and $K = 5$ when N = 445.

Figures 1.15 and 1.16 show a PDF graph and $S(K)$ plot for a simulated sample size of N = 150. The true optimal K value is $K = 1$. However, the mode of the PDF of optimal K value is at $K = 0$. It appears as if the procedure is selecting a slightly sub-optimal K. This is because the true $S(0)$, $S(1)$, and $S(2)$ are all within .7% of each other. The frequency of the $K = 0$ estimator is about .47 while the frequency of the

$K=1$ estimator is about .3. Each is extremely close to being optimal. Hence, if two K values are very close, the $\hat{S}(K)$ criterion seems to select the smaller one. Figures 1.17 and 1.18 show the same types of plots, except $(\hat{\beta}_2, \hat{\beta}_3)$ has been doubled. The intuition behind this is that by making X stronger in the outcome equation, the $K=1$ estimator should become relatively more efficient than the $K=0$ estimator and the PDF plot should reflect this. In fact, it does. The PDF is now tighter and the mode is at $K=1$ with a frequency of approximately .5.

Figures 1.19 and 1.20 show the two plots when N = 445. Remember, when N = 25 the true optimal $K=0$ and when N = 150 the true optimal $K=1$, but marginally. In increasing sample size to 445, one would expect the true optimal K to more clearly be $K=1$. The graphs show this to be true. There is now a 1.1% difference between the true $S(0)$ and $S(1)$. There still exists a significant selection of $K=0$, but the mode of the PDF is at $K=1$ with a frequency of about .45 as compared to a frequency of about .22 for $K=0$. In Figures 1.21 and 1.22 the sample size remains at 445, but $(\hat{\beta}_2, \hat{\beta}_3)$ has been doubled again. Now there exists a 2.7% difference between the true $S(0)$ and $S(1)$. One can see that this change drastically tightens up the left tail of the PDF. $K=1$ is now chosen 65% of the time and is the overwhelmingly clear choice from the PDF plot.

Finally, instead of calibrating a data generating process to the LaLonde dataset, the results of the procedure on the dataset itself are examined. The known propensity score is assumed to be $p = \frac{1}{N}\sum_{i=1}^{N} t_{Li} = .416$[19]. Table 1.3 presents the results of the GMM procedure. All estimates of the average treatment effect are in a similar range. The $K = -1$ estimator and the $K = 0$ estimator are identical since the true propensity score is assumed to equal the sample propensity score. According to the GMM procedure, the $K = 1$ estimator is higher-order optimal, $\hat{K} = 1$. Thus, the standard difference-in-averages estimator is sub-optimal. There is a 2.6% difference in the $K = 1$ estimator relative to the more common difference-in-averages, $K = 0$, estimator. Depending on the purpose of the study and the use of the results the 2% – 6% difference between the set of possible estimators may or may not be significant.

This section essentially reinforced many of the results and much of the intuition of Section 1.6. It did so in the context of a more realistic fake data generating process. One can again see that the stronger X is in the outcome equation, the higher the optimal K. The larger the sample size, the higher the optimal K. The criterion tends to select the lower of two K values when one is optimal and the other is very close to optimal. Finally, these simulations show that moment selection based on the $\hat{S}(K)$ criterion works well for the models considered and it appears to work well for the actual experimental LaLonde dataset.

[19] The results are very similar when the true propensity score is assumed to equal .5.

1.8 Conclusions

This chapter started by laying out the fundamental problem with respect to estimation and analysis of average treatment effects. It reviewed previous work, based on asymptotic theory, which states that estimation of average treatment effects with the true propensity score is unbiased but inefficient while estimation of average treatment effects with the non-parametric propensity score is consistent and asymptotically efficient. Since these estimators would tend to be sub-optimal in finite samples, a GMM model is considered in which a finite amount of auxiliary information reflecting one's knowledge of the propensity score is used to improve efficiency. MSE plots from this model provided intuition for the certain types of situations in which this efficiency gain may be significant and when it may not. In particular, the true propensity score estimator always performed extremely poorly. There should be a non-trivial efficiency gain relative to the difference-in-averages estimator when X is strong in the outcome equation. Also, there should be a non-trivial efficiency gain relative to an estimator similar to the HIR non-parametric propensity score estimator when sample size is small.

To develop an actual procedure that can be implemented in practice, asymptotic expansions of the average treatment effect estimator and the Lagrange Multiplier estimator were derived. Through these expansions, a higher-order approximation to the true MSE was developed. It was shown, through simulations, that this higher-order formula yields a tight approximation to the true MSE and that the second order terms are important in yielding this tight fit. Taking the components of this formula that depend on K, a higher-order moment selection criterion, $\hat{S}(K)$, was developed. It was proven, in a fashion similar to Donald and Newey (1999) and Li (1987), that the selection rule for $\hat{S}(K)$ is higher-order asymptotically optimal for $S(K)$.

Finally, the finite sample performance of the selection rule was tested on simulated data from an arbitrary data generating process and on simulated data from a data generating process calibrated to LaLonde's 1986 dataset. The results tended to reinforce the previously developed intuition. It also appears that the selection rule selects the lower K value when two adjacent K values are very close to optimal, even if the larger K value is the optimal one. The PDF of the optimal K value over 1,000 iterations showed that the selection rule tends to pick the optimal K, or a value very close to it MSE-wise, the majority of the time.

The results of this chapter assume that the propensity score is constant and known, and that X is a scalar. Chapter 2 extends the analysis of this chapter by examining the case where X is a vector. The analysis of these chapters has the most value for research in an experimental setting. Lastly, Chapter 3 extends the analysis of the previous chapters by considering the case when the propensity score is not constant and is unknown. Examination of this case is valuable because this holds in the case of a non-experimental setting and analysis of non-experimental data is becoming more frequently observed in economics.

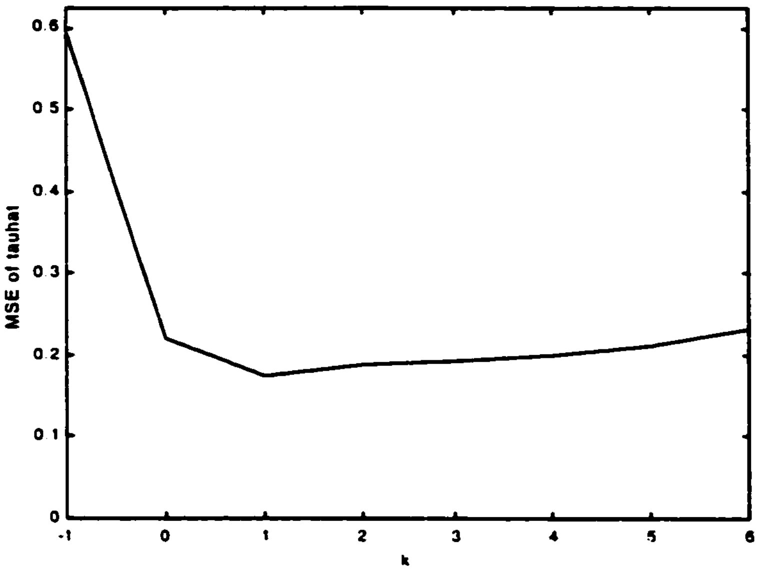

Figure 1.1: True MSE of $\hat{\tau}$ ($y = .5 + 2t + x + \varepsilon$, p(x) = 1/2, N = 25)

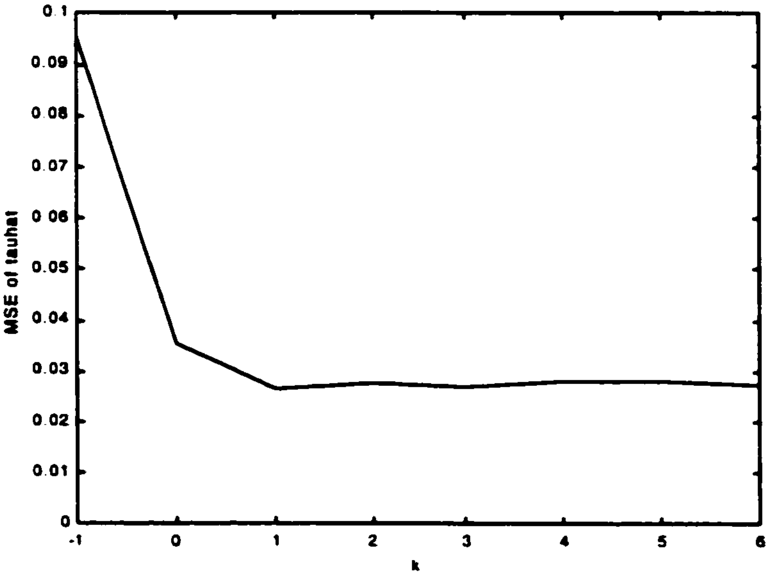

Figure 1.2: True MSE of $\hat{\tau}$ ($y = .5 + 2t + x + \varepsilon$, p(x) = 1/2, N = 150)

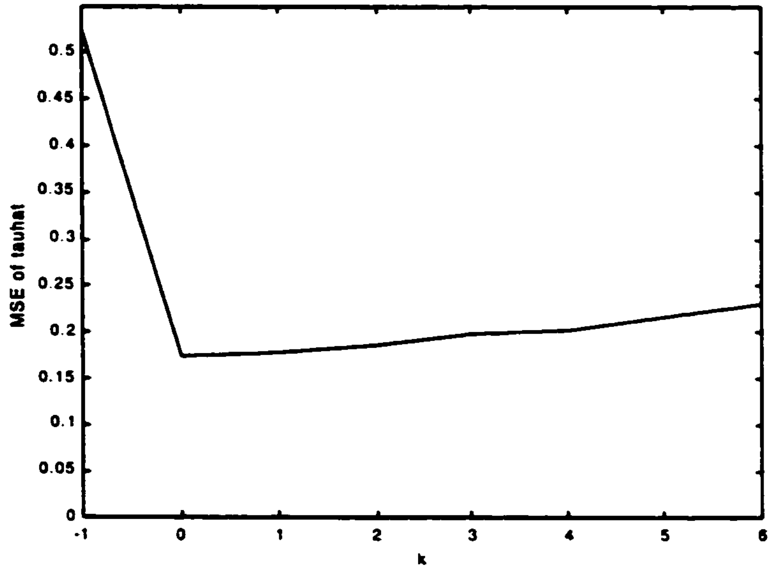

Figure 1.3: True MSE of $\hat{\tau}$ ($y = .5 + 2t + .3x + \varepsilon$, p(x) = 1/2, N = 25)

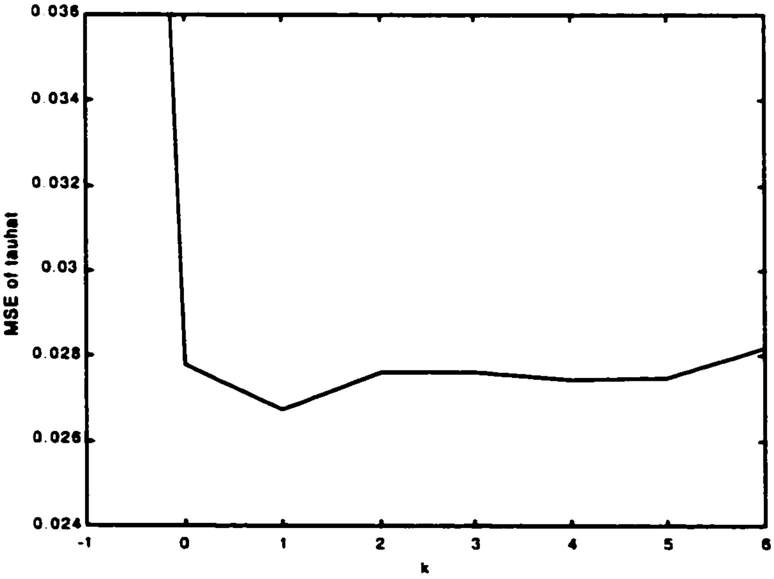

Figure 1.4: True MSE of $\hat{\tau}$ ($y = .5 + 2t + .3x + \varepsilon$, p(x) = 1/2, N = 150) (Zoomed In)

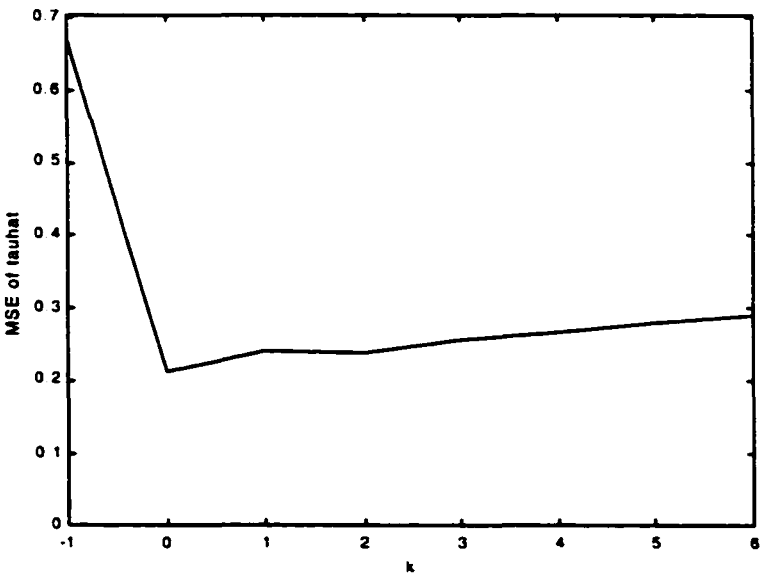

Figure 1.5: True MSE of $\hat{\tau}$ ($y = .5 + 2t + x + \varepsilon$, p(x) = Logit, N = 25)

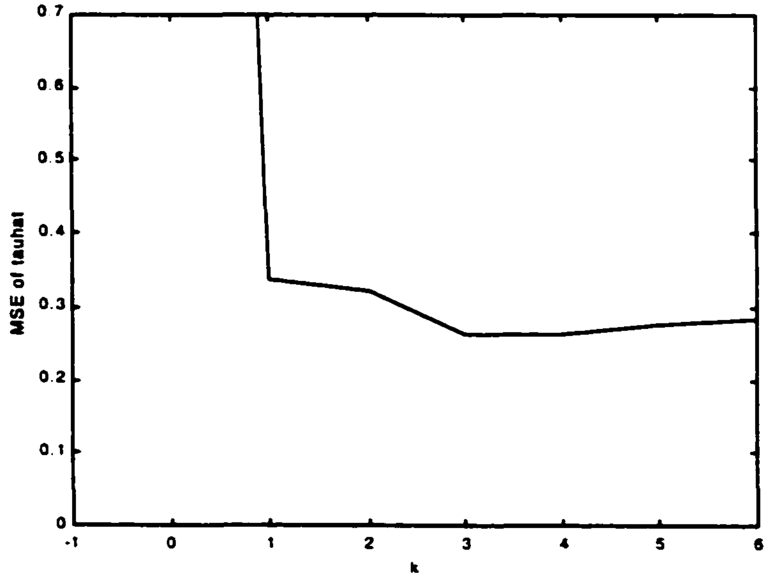

Figure 1.6: True MSE of $\hat{\tau}$ ($y = .5 + 2t + 8x + \varepsilon$, p(x) = Logit, N = 25)

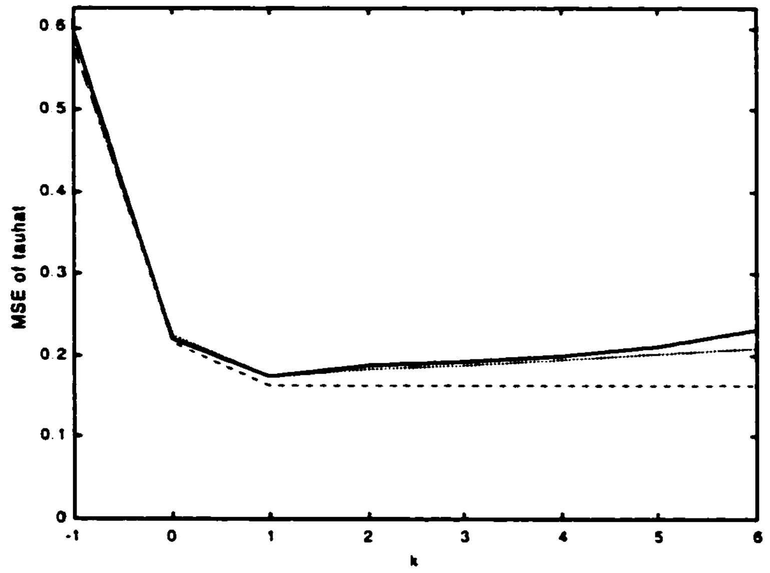

Figure 1.7: True MSE (Solid), Formula MSE (Dotted), and 1st Order Formula MSE (Dash-Dot)

($y = .5 + 2t + \ x + \varepsilon$, p(x) = 1/2, N = 25)

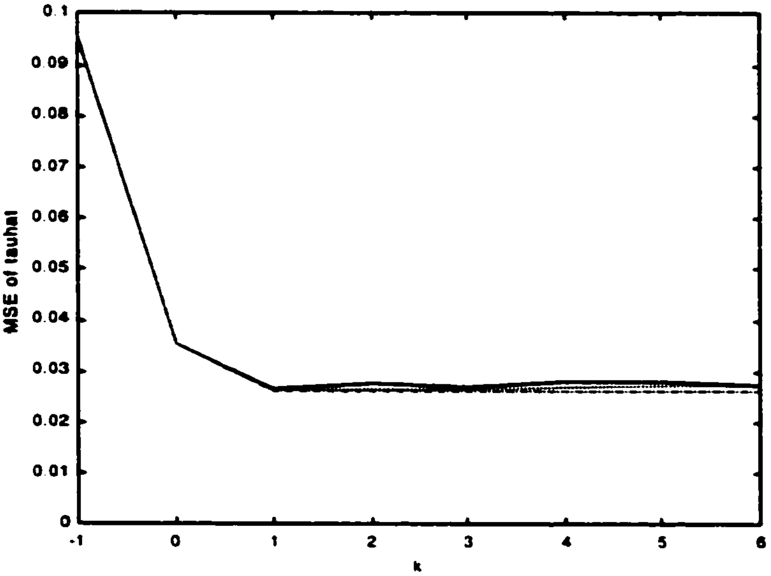

Figure 1.8: True MSE (Solid), Formula MSE (Dotted), and 1st Order Formula MSE (Dash-Dot)

($y = .5 + 2t + \ x + \varepsilon$, p(x) = 1/2, N = 150)

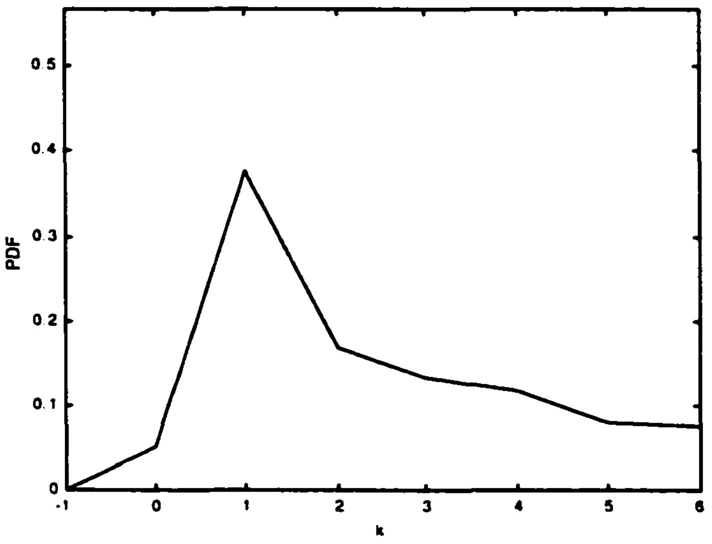

Figure 1.9: PDF of $\hat{K}$ Over 1000 Simulations

($y = .5 + 2t + x + .4tx + \varepsilon$, p(x) = 1/2, N = 25)

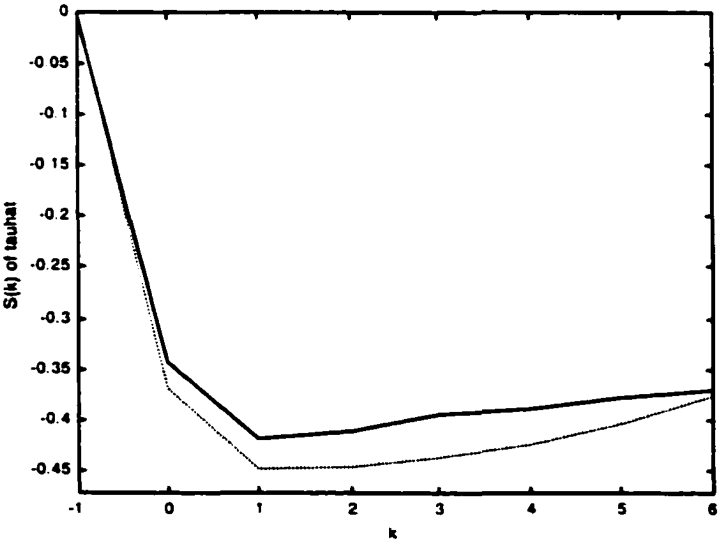

Figure 1.10: True $S(K)$ (Solid) and Mean of $\hat{S}(K)$ (Dotted) Over 1000 Simulations

($y = .5 + 2t + x + .4tx + \varepsilon$, p(x) = 1/2, N = 25)

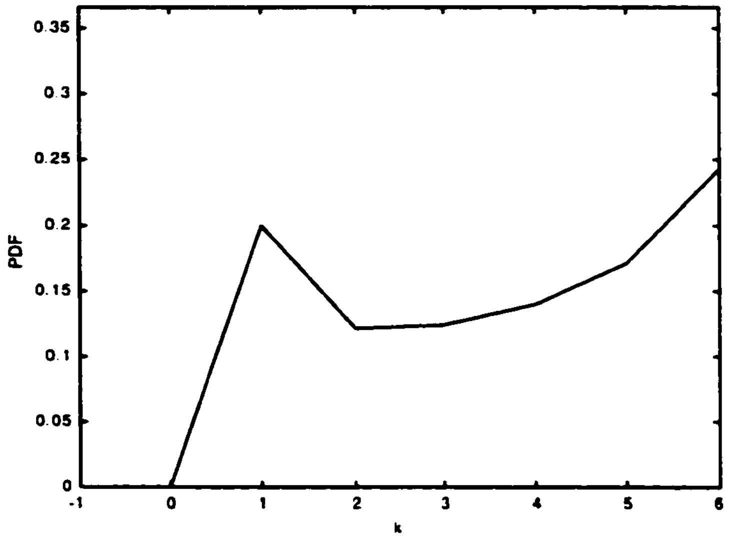

Figure 1.11: PDF of $\hat{K}$ Over 1000 Simulations

($y = .5 + 2t + x + .4tx + \varepsilon$, p(x) = 1/2, N = 150)

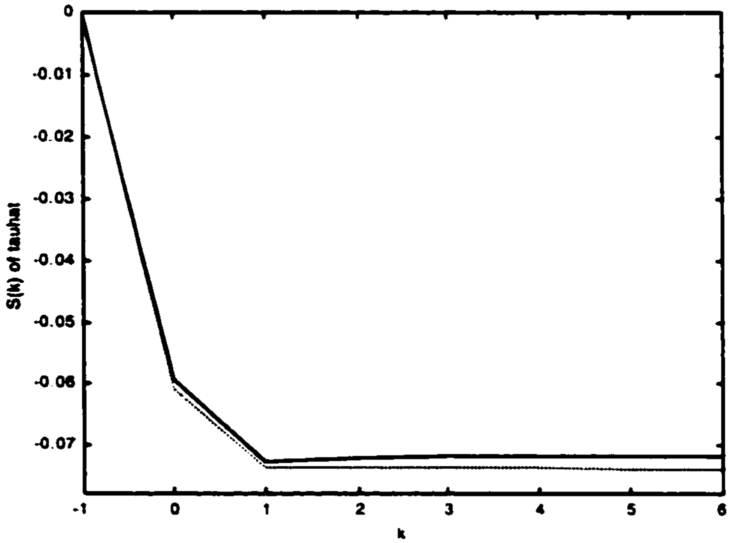

Figure 1.12: True $S(K)$ (Solid) and Mean of $\hat{S}(K)$ (Dotted) Over 1000 Simulations

($y = .5 + 2t + x + .4tx + \varepsilon$, p(x) = 1/2, N = 150)

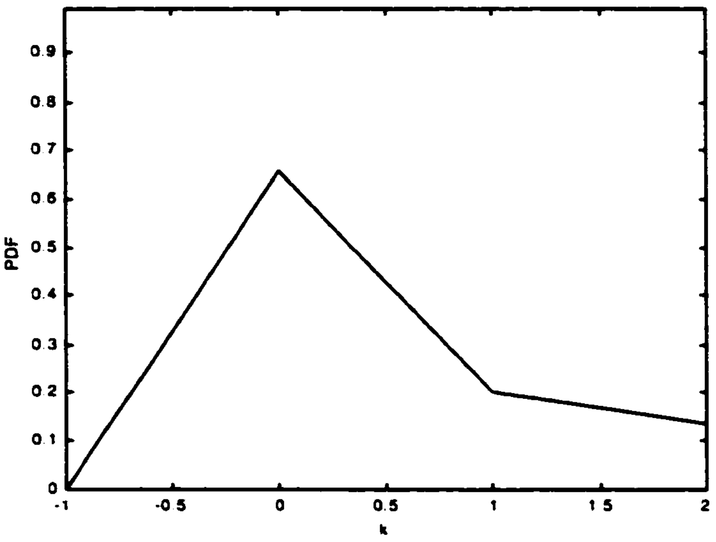

Figure 1.13: PDF of $\hat{K}$ for LaLonde Calibrated DGP

($y = 4359 + 1711t + .155x + .028tx + \varepsilon$, p(x) = .416, N = 25)

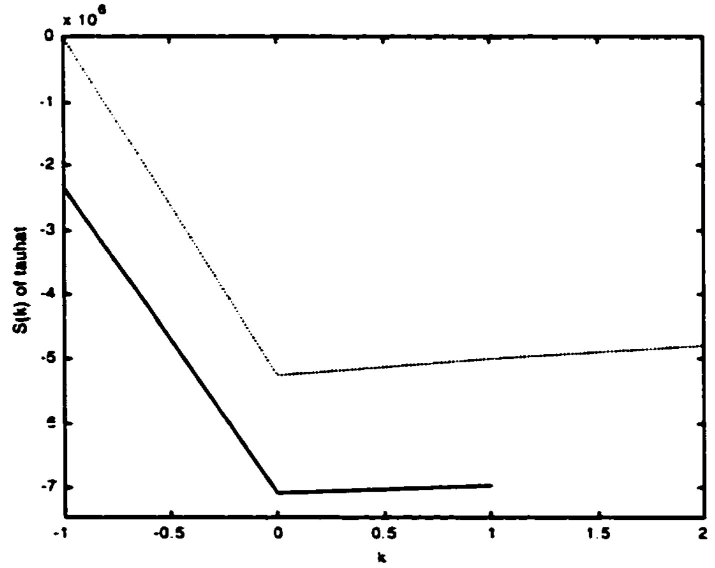

Figure 1.14: True $S(K)$ (Solid) and Mean of $\hat{S}(K)$ (Dotted) for LaLonde Calibrated DGP

($y = 4359 + 1711t + .155x + .028tx + \varepsilon$, p(x) = .416, N = 25)

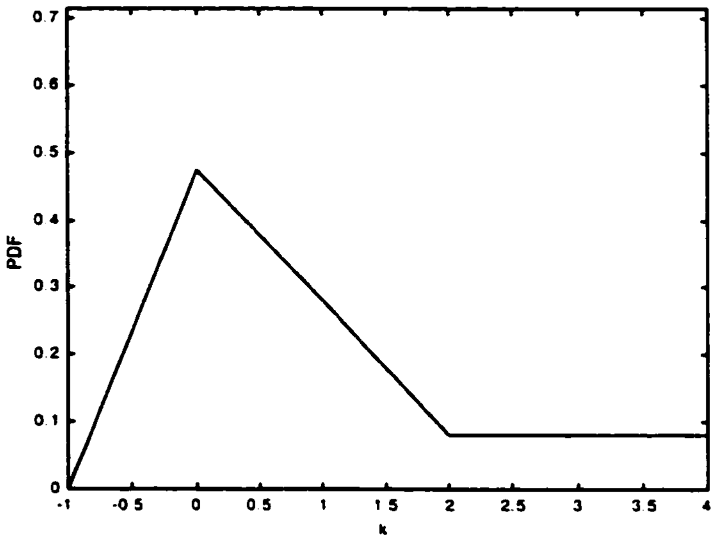

Figure 1.15: PDF of $\hat{K}$ for LaLonde Calibrated DGP

($y = 4359 + 1711t + .155x + .028tx + \varepsilon$, p(x) = .416, N = 150)

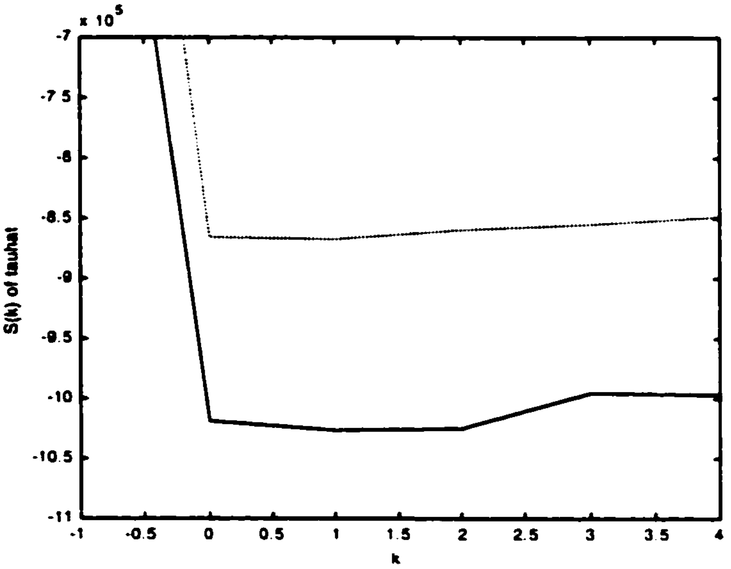

Figure 1.16: True $S(K)$ (Solid) and Mean of $\hat{S}(K)$ (Dotted) for LaLonde Calibrated DGP

($y = 4359 + 1711t + .155x + .028tx + \varepsilon$, p(x) = .416, N = 150)

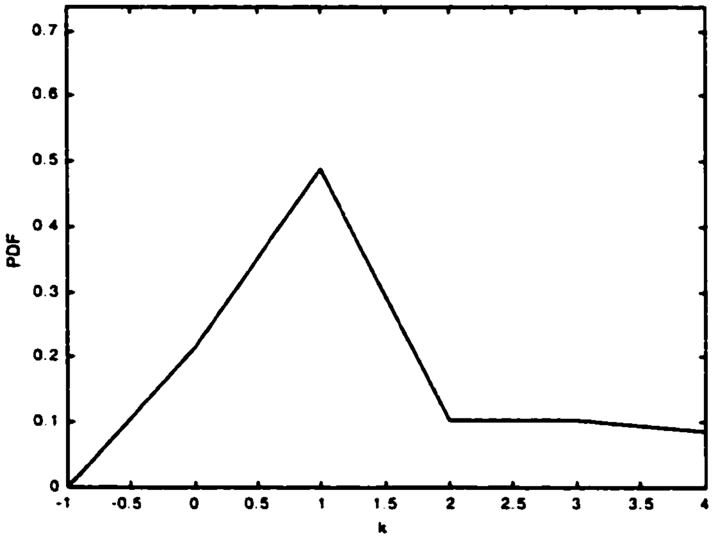

Figure 1.17: PDF of $\hat{K}$ for LaLonde Calibrated DGP

($y = 4359 + 1711t + .310x + .055tx + \varepsilon$, p(x) = .416, N = 150)

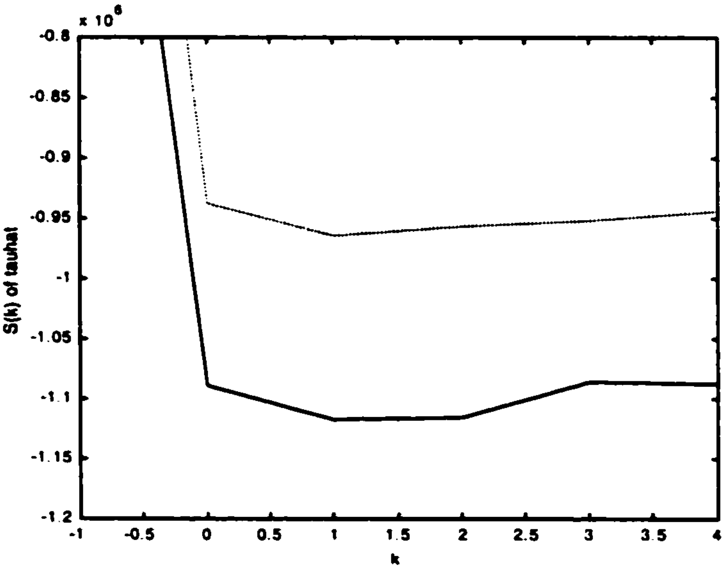

Figure 1.18: True $S(K)$ (Solid) and Mean of $\hat{S}(K)$ (Dotted) for LaLonde Calibrated DGP

($y = 4359 + 1711t + .310x + .055tx + \varepsilon$, p(x) = .416, N = 150)

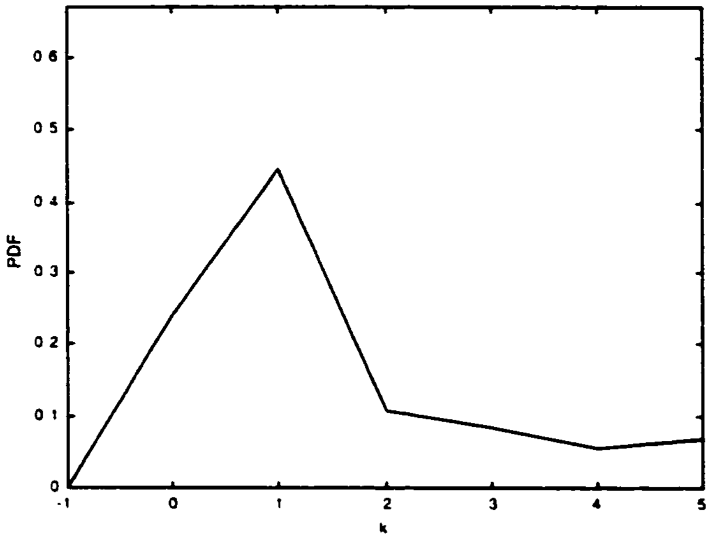

Figure 1.19: PDF of $\hat{K}$ for LaLonde Calibrated DGP

($y = 4359 + 1711t + .155x + .028tx + \varepsilon$, p(x) = .416, N = 445)

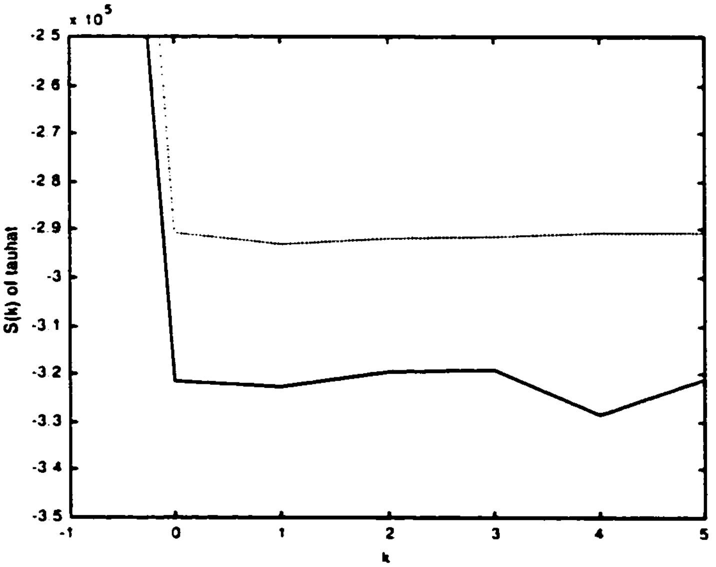

Figure 1.20: True $S(K)$ (Solid) and Mean of $\hat{S}(K)$ (Dotted) for LaLonde Calibrated DGP

($y = 4359 + 1711t + .155x + .028tx + \varepsilon$, p(x) = .416, N = 445)

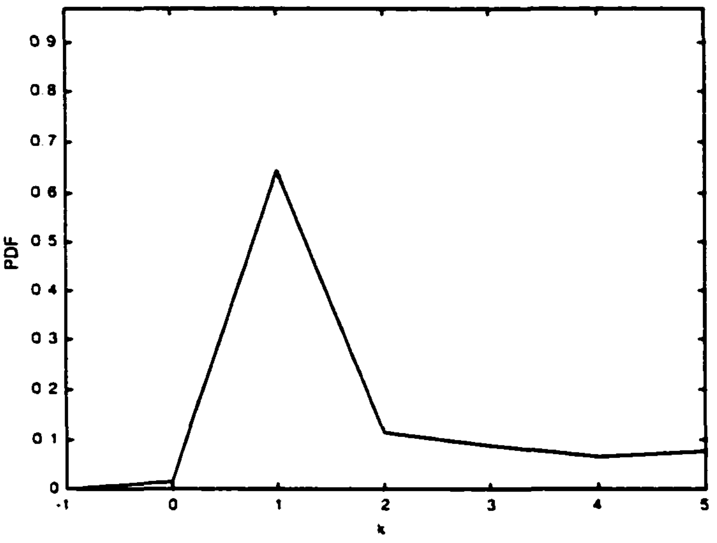

Figure 1.21: PDF of $\hat{K}$ for LaLonde Calibrated DGP

($y = 4359 + 1711t + .310x + .055tx + \varepsilon$, p(x) = .416, N = 445)

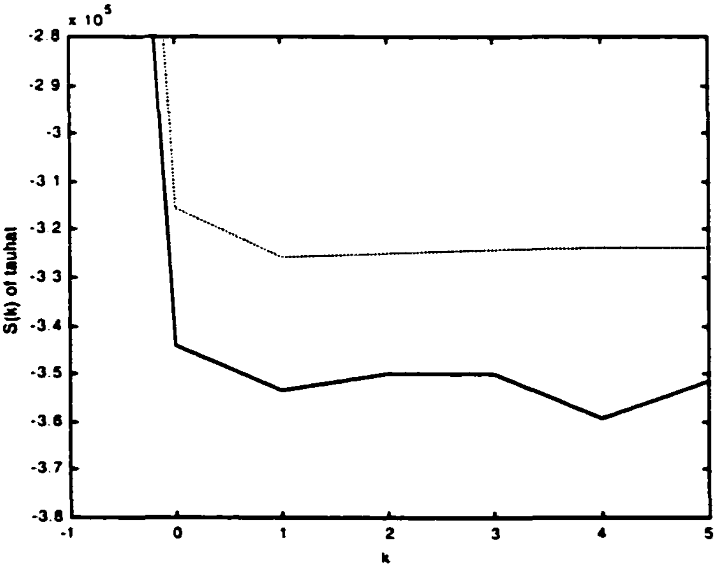

Figure 1.22: True $S(K)$ (Solid) and Mean of $\hat{S}(K)$ (Dotted) for LaLonde Calibrated DGP

($y = 4359 + 1711t + .310x + .055tx + \varepsilon$, p(x) = .416, N = 445)

	Full Sample	Treated	Controls
Earnings 1978 (Y)	5300 (6632)	6349 (7867)	4555 (5484)
Treatment Status (T)	.416 (.493)	1 (0)	0 (0)
Earnings 1975	1377 (3151)	1532 (3219)	1267 (3103)
Age	25.4 (7.1)	25.8 (7.2)	25.1 (7.1)
Education	10.2 (1.8)	10.4 (2.0)	10.1 (1.6)
Unemployed 1975	.649 (.478)	.600 (.491)	.685 (.466)
Married	.169 (.375)	.189 (.393)	.154 (.362)
Black	.834 (.373)	.843 (.365)	.827 (.379)
Hispanic	.088 (.283)	.060 (.237)	.108 (.311)
No Degree	.782 (.413)	.708 (.456)	.835 (.372)

Table 1.1: Sample Means (Standard Deviation) of Post-Training Earnings, Treatment Status, and Pre-Treatment Variables for a Sub-Sample of the NSW AFDC Participants of LaLonde (1986).

	Non-Standardized	Standardized
Constant	1545	5301
Treatment Status (T)	1610 [2.52]	793.3 [2.52]
Earnings 1975	.101 [.786]	316.8 [.786]
Age	55.9 [1.23]	396.6 [1.23]
Education	369.7 [1.62]	661.7 [1.62]
Unemployed 1975	-499.6 [-.604]	-238.4 [-.604]
Married	-134.9 [-.153]	-50.5 [-.153]
Black	-2108 [-1.80]	-784.8 [-1.80]
Hispanic	137.1 [.088]	38.8 [.088]
No Degree	-187.5 [-.188]	-77.4 [-.188]

Table 1.2: Non-Standardized and Standardized OLS Results for a Regression of Earnings in 1978 on Treatment Status and Other Pre-Treatment Variables for a Sub-Sample of the NSW AFDC Participants of LaLonde (1986). T-Statistics in Brackets.

K	Average Treatment Effect Estimate
-1	1794.3
0	1794.3
_1_*	_1748.9_
2	1726.4
3	1844.8
4	1838.4
5	1809.7

Table 1.3: Average Treatment Effect Estimates for a Sub-Sample of the NSW AFDC Participants of LaLonde (1986). One Covariate: Earnings in 1975. *Optimal K = 1.

2. Chapter 2

Higher-Order Optimal Estimation of Binary Average Treatment Effects with Experimental Data and a Vector of Covariates

2.1 Introduction

In experimental settings, the econometrician will most often have information on a vector of pre-treatment variables rather than just a scalar. Even though there is no selection problem in this setting, the results of Chapter 1 imply that information from the vector of covariates may have a significant value. In particular, if a covariate is relatively strong in the outcome equation, the relative value of inclusion of moments of this covariate will be high. As has been shown with one covariate, there can be a significant MSE gain from using these covariates to account for sample divergences from true distributional probabilities.

This chapter will show that the results and intuition of Chapter 1 still hold in a setting with a vector of covariates. It will also show that the higher-order asymptotically optimal moment selection procedure of Chapter 1 is a valid and useful tool for practitioners in this new setting. The econometrician's knowledge of the propensity score and of a pre-treatment covariate vector can lead to a non-trivial MSE reduction relative to the true propensity score estimator, the difference-in-averages estimator, and an estimator similar to the HIR non-parametric propensity score estimator. In this chapter, it is assumed that the econometrician's decision problem involves choosing the optimal *polynomial order* of the vector of covariates for inclusion. Thus, the econometrician has already ex-ante selected the best subset of covariates to include in the overall estimation procedure. In practice, this may be accomplished through theoretical justifications, ex-ante regressions of the outcome on the treatment indicator and the covariate vector, or a combination of the two. Given this setup, the theoretical results of Chapter 1 still hold. Theorems 1.1 and 1.2 can then be used to construct a higher-order moment selection criterion and this selection criterion can be implemented in practice in a fashion similar to Chapter 1.

The remainder of the chapter is as follows. Section 2.2 discusses the setup of the new model and shows that the fundamental results of Chapter 1 remain unchanged. Section 2.3 presents results of Monte Carlo simulations with two covariates under a variety of artificial data generating processes. Section 2.4 presents results of Monte Carlo simulations based on data generating processes calibrated to a subset of the dataset of LaLonde (1986) with "Earnings in 1975" and "Education" as the two covariates. Finally, Section 2.5 concludes.

2.2 Setup

The setup of the model is identical to Section 1.2 except X is now assumed to be a $(D \times 1)$ vector rather than a scalar. Thus, the average treatment effect is estimated in a new GMM system of moments. In particular, let

$$\bar{\psi}(y,t,x,\tau) = \begin{pmatrix} \bar{\psi}_1(y,t,x,\tau) \\ \bar{\psi}_2(t,x) \end{pmatrix},$$

where

$$\bar{\psi}_1(y,t,x,\tau) = \left(\frac{y \cdot t}{p(x)} - \frac{y \cdot (1-t)}{1-p(x)} \right) - \tau = \bar{V} - \tau,$$

$$\bar{\psi}_2(t,x) = \begin{pmatrix} t - p(x) \\ f_1(x)(t-p(x)) \\ \vdots \\ f_P(x)(t-p(x)) \end{pmatrix},$$

$$E(\bar{\psi}(y,t,x,\tau)) = 0,$$

where $p(x)$ is the true propensity score. Notice the only change is that x^j has been replaced by $f_j(x)$ in the $\bar{\psi}_2$ equation. This change is due to the fact that higher-order polynomials and interactions of the elements of X may be included in the estimation procedure as opposed to just higher-order polynomials of a scalar variable.

THEOREM 2.1: *Suppose X is a vector and Assumptions 1 – 5 hold. Asymptotic Expansions 1.1 and 1.2, Lemmas 1.1 and 1.2, Corollary 1.1, and Theorems 1.1 and 1.2 are all valid.*

Proof: See Appendix.

Theorem 2.1 shows that the asymptotic expansions of Chapter 1 remain correct. Also, the higher-order MSE approximation and, hence, the higher-order $S(K)$ approximation remain correct. Finally, the criterion function $\hat{S}(K)$ remains higher-order asymptotically optimal for $S(K)$. Given this result, one can employ the moment selection procedure of Chapter 1 on the model above and obtain a higher-order asymptotically optimal average treatment effect estimate.

The key procedural assumption of this chapter is as follows. X is the subset of covariates upon which the econometrician has ex-ante chosen to base estimation. The econometrician decides whether or not to include all first-order terms. In other words, one decides whether or not to include the $D\times 1$ vector $f_1(x)=x$. If this vector is included, one then decides whether or not to include all second-order terms, $f_2(x)=x^2$. This continues until the optimal number of moments is chosen, $K^* = D\cdot P^*$, where P^* is the optimal polynomial order. This assumption circumvents the ordering problem associated with a vector of covariates. If the econometrician could effectively ex-ante order the covariates, marginal first-order moments could be included based on this ordering. Then, if MSE could still be reduced, marginal second-order moments could be included in the same fashion and so on. The benefit of this method is that inclusion of marginal moments can be done on a covariate-by-covariate basis rather than an order-by-order basis[20]. It will be shown below that inclusion of moments on a covariate-by-covariate basis could lead to an efficiency gain, relative to inclusion on an order-by-order basis, in certain instances.

2.3 Monte Carlo Simulations with Two Covariates

This section presents results based on Monte Carlo simulations under a variety of artificial data generating processes. Similar to Section 1.6, it provides evidence of the performance of $\hat{S}(K)$ and the GMM procedure with actual finite sample datasets. As in Chapter 1, the procedure is executed for 1,000 different Monte Carlo samples. The distribution of the optimal *order* over these 1,000 samples is plotted and analyzed. Also, the mean of $\hat{S}(K)$ over the 1,000 samples is plotted and compared to the true $S(K)$.

Let

$$Y_i = .5 + 2T_i + \beta_2 X_{1i} + \beta_3 X_{2i} + \beta_4 X_{1i}^2 + \beta_5 X_{2i}^2 + .4T_i X_{1i} + .6T_i X_{2i} + \varepsilon_i,$$

$$\varepsilon_i \sim N(0,1),$$

$$X_i \sim UNIF(-1,1),$$

$$p(X_i) = \tfrac{1}{2}.$$

In the explanation of the results that follows, one figure will show the PDF of the optimal order, P^*, from 1000 simulations. Another figure will show the true $S(K)$ and the mean of $\hat{S}(K)$ over 1000 simulations. The PDF graph yields results on an order-by order basis

[20] The results will show that there can be a MSE gain associated with ordering. Knowledge of an acceptable ordering of X can be valuable.

in order to see how well the procedure works for selection of the optimal order. The last order value is always from an interaction marginal moment restriction, just to see if this has an effect. The $S(K)$ graph yields results on a moment-by-moment basis. The $K=1$ estimator includes the linear term of X_1 as a marginal moment restriction. The $K=2$ estimator includes the linear term of X_2 as a marginal moment restriction. The $K=3$ and $K=4$ estimators work in the same way except with second-order terms of X_1 and X_2. Finally, the last K value includes an interaction marginal moment restriction. This method of displaying results was done intentionally because any order-by-order $S(P)$ values are also contained in moment-by-moment $S(K)$ results, however in certain situations the moment-by-moment results will show a potential MSE gain from using a moment-by-moment procedure with a pre-ordering of X.

Figures 2.1 and 2.2 show results for the case where $(\beta_2,\beta_3,\beta_4,\beta_5)' = (1,1,0,0)'$ and $N=25$. Figure 2.2 shows that the optimal K value is $K=2$. This estimator includes all first-order terms of X. The $K=2$ estimator contains a 76.8% MSE reduction[21] relative to the true propensity score $K=-1$ estimator, a 38.9% MSE reduction relative to the difference-in-averages $K=0$ estimator, and a 15.9% MSE reduction relative to an estimator similar to the HIR non-parametric propensity score $K=5$ estimator. Accordingly, Figure 2.1 shows that the PDF of the optimal polynomial order has its mode at $P=1$ with a frequency of close to .6. Also, the right tail of the

[21] MSE values not shown in the graphs.

distribution tends to be thicker due to the fact that the true $S(K)$ is closer to the optimal $S(K)$ for K values greater than $K = 2$ versus those less than $K = 2$.

Figures 2.3 and 2.4 show results for the case where $(\beta_2, \beta_3, \beta_4, \beta_5)' = (1,0,0,1)'$ and $N = 25$. Figure 2.4 shows that the optimal K value is $K = 2$. Again, this estimator includes all first-order terms of X. The optimal $K = 2$ estimator contains a 75.3% MSE reduction relative to the true propensity score $K = -1$ estimator, a 30.7% MSE reduction relative to the difference-in-averages $K = 0$ estimator, and a 25.1% MSE reduction relative to an estimator similar to the HIR non-parametric propensity score $K = 7$ estimator. Figure 2.3 shows that the PDF of the optimal polynomial order has its mode at $P = 1$ with a frequency of close to .5. Again, the right tail of the distribution tends to be thicker due to the fact that the true $S(K)$ is closer to the optimal $S(K)$ for K values greater than $K = 2$ versus those less than $K = 2$. Notice that X_2^2 enters the outcome equation and X_2 doesn't. These results show that, with this data generating process, inclusion of the linear term of X_2 is optimal in accounting for X_2^2 in the outcome equation. The cost of including all second-order terms is greater than the benefit even though the second-order terms include X_2^2. Thus, X_2^2 is sufficiently strong in the outcome equation such that it is optimal to include X_2 in the estimation procedure, but it is not strong enough such that X_2^2 should also be included.

Figures 2.5 and 2.6 show results for the same β values as Figures 2.3 and 2.4, except $N = 150$. Figure 2.6 shows that the optimal K value is still $K = 2$. Now, the optimal $K = 2$ estimator contains a 77.4% MSE reduction relative to the true propensity score $K = -1$ estimator, a 33.7% MSE reduction relative to the difference-in-averages $K = 0$ estimator, and a 4.0% MSE reduction relative to an estimator similar to the HIR non-parametric propensity score $K = 9$ estimator. The drastic reduction in the efficiency gain of the optimal estimator relative to the large K estimator is due to the increased sample size. As would be expected, Figure 2.5 shows that the PDF of the optimal polynomial order is zero for $P = -1$ and $P = 0$ and is positive and significant for $P \geq 1$. Because the sample size is large, the average treatment effect estimator is essentially the same, in a MSE sense, when higher-order terms of X are included in the estimation procedure. This is because the cost of additional noise has decreased.

Figures 2.7 and 2.8 show results for the case where $(\beta_2, \beta_3, \beta_4, \beta_5)' = (1,1,0,7)'$ and $N = 25$. Figure 2.8 shows that the optimal K value is $K = 4$. This estimator includes all first-order and second-order terms of X. The optimal $K = 4$ estimator contains a 91.5% MSE reduction relative to the true propensity score $K = -1$ estimator, a 76.1% MSE reduction relative to the difference-in-averages $K = 0$ estimator, and a 16.1% MSE reduction relative to an estimator similar to the HIR non-parametric propensity score $K = 7$ estimator. In this specification, X_1 is slightly strong in the outcome equation and X_2 is very strong in the outcome equation. Figure 2.8 shows that $S(3) > S(2)$. In other words, MSE rises when the marginal moment of X_1^2 is included along with the first-order moments. However, since X_2 is strong in the outcome equation, $S(4) < S(2) < S(3)$. The MSE is reduced when the marginal moment of X_2^2

is included along with the moment from X_1^2 and all first-order moments. Thus, if the econometrician could effectively order X_1 and X_2, the most optimal estimator would most likely be the estimator employing all first-order moments and the X_2^2 moment. Finally, Figure 2.7 shows that the PDF of the optimal polynomial order correctly has its mode at $P = 2$ with a frequency of close to .6.

The results of this section reaffirm the results and intuition of Chapter 1. Again it is shown that there can be a significant MSE gain relative to the true propensity score estimator and the difference-in-averages estimator when X is strong in the outcome equation. Also, there can be a significant MSE gain relative to an estimator similar to the HIR non-parametric propensity score estimator when sample size is small. This section has shown that these results also hold true when X is a vector. In relevant small sample sizes, the procedure correctly picked the optimal order of X to include. By "correct" it is meant that the mode of the PDF was at the correct order value and the PDF value for that order was high. Finally, it was also shown that, in certain situations, there might be a possible MSE gain from pre-ordering the variables of X and implementing the procedure on a moment-by-moment basis rather than an order-by-order basis.

2.4 Application to Experimental Data of LaLonde (1986)

In this section, the GMM procedure is applied to various calibrated data generating processes of a subset of the LaLonde (1986) experimental dataset. Then, the procedure is applied to the actual dataset itself. The data generating process is calibrated to the empirical distribution of the relevant variables in this dataset. The two covariates used in this section are "Earnings in 1975" and "Education." In doing a study on future earnings, it seems reasonable to take into account the current earnings of the participants. Education was chosen as the second variable because it is the most significant non-binary[22] covariate in Table 1.2. The results will reinforce all previous results and show that the proposed estimation procedure performs well with a vector of covariates and a more credible data generating process.

Similar to Chapter 1, a Monte Carlo dataset is calibrated to the LaLonde dataset according to the following rules. Let the scalar covariate x_{L1} be Earnings in 1975 and let x_{L2} be Education. Let (y_L, t_L) be the outcome vector and the treatment vector from LaLonde's dataset, respectively. Draw (x_1, x_2) from the empirical joint distribution of (x_{L1}, x_{L2}). Let the propensity score equal the empirical mean of the treatment vector t_L, $p = \frac{1}{N}\sum_{i=1}^{N} t_{Li} = .416$. Draw t from a $BIN(1,p)$ distribution. Run the OLS regression of

$$y_L = \beta_0 + \beta_1 t_L + \beta_2 x_{L1} + \beta_3 x_{L2} + \beta_4 t_L x_{L1} + \beta_5 t_L x_{L2} + \varepsilon .$$

Let $\hat{\beta}$ be the resulting vector of parameter estimates and $\hat{\sigma}^2 = \hat{\varepsilon}'\hat{\varepsilon} / (N-6)$. Draw v from a $N(0, \hat{\sigma}^2)$ distribution. Finally, let

$$y = \hat{\beta}_0 + \hat{\beta}_1 t + \hat{\beta}_2 x_1 + \hat{\beta}_3 x_2 + \hat{\beta}_4 t x_1 + \hat{\beta}_5 t x_2 + v .$$

In what follows, this calibration routine is used to examine the properties of the $\hat{S}(K)$ criterion. The simulated sample size is adjusted from N = 25 to N = 150 to N = 445

[22] A non-binary covariate is chosen so that inclusion of higher-order polynomials may matter.

(LaLonde's sample size), and the $\left(\hat{\beta}_2, \hat{\beta}_3, \hat{\beta}_4, \hat{\beta}_5\right)$ values are adjusted to see how they affect the results.

Figures 2.9 and 2.10 show the results of the calibrated model when $N = 25$. The optimal estimator is the $K = 0$ estimator. The true $S(1)$ is close to $S(0)$. However, the true $S(2)$ is significantly different from $S(0)$. Since the vector procedure chooses the optimal order, P^*, rather than the optimal moments, K^*, the PDF of the optimal order is relatively tightly fit around $P = 1$. Notice that the true $S(K)$ function disappears after $K = 2$. Again, this is solely because the Earnings in 1975 variable is 65% zeros. If one of the 1,000 fake datasets has an x_1 with a large amount of zeros, then x_1 will be close to a linear combination of its powers and the routine will fail or return NaNs with high K values. For $N = 150$ and $N = 445$, this is not a problem. Figures 2.11 and 2.12 show the results of the same model when $\left(\hat{\beta}_2, \hat{\beta}_3, \hat{\beta}_4, \hat{\beta}_5\right)$ has been doubled in magnitude. The optimal estimator is the $K = 1$ estimator and the optimal order is the $P = 0$ order. In this case, the true $S(1)$ is now less than $S(0)$ and the true $S(2)$ is much closer to $S(0)$. The effect of this is that PDF of the optimal order has a mode at $P = 0$ of .45, but the PDF value of $P = 1$ is high too at .33. Also, the distribution of the PDF has a thicker right tail than the PDF of Figure 2.9. These results are as one would expect them to be.

Figures 2.13 and 2.14 show the results of the calibrated model when $N = 150$. At $N = 25$, the optimal estimator was the $K = 0$ estimator. Now, the optimal estimator increases to the $K = 2$ estimator where all first order terms are included. Since the true

$S(2)$ is relatively close to $S(0)$ in percentage terms, the PDF of the optimal order has its mode at $P=1$ with a frequency of .45, but the PDF value at $P=0$ is also relatively high at .33. Figures 2.15 and 2.16 show the results of the same model when $(\hat{\beta}_2, \hat{\beta}_3, \hat{\beta}_4, \hat{\beta}_5)$ has been doubled in magnitude. The optimal estimator is still the $K=2$ estimator and the optimal order is the $P=1$ order. As would be expected, the PDF of the optimal order still has its mode at $P=1$ with a higher frequency of .62 and the distribution has become much more tight on the left. The procedure picks a low order with much less frequency in this case. This is due to the face that the MSE penalty of doing so is significantly greater, in percentage terms, because X_1 and X_2 have been made stronger in the outcome equation.

Figures 2.17 and 2.18 show the results of the calibrated model when $N=445$. The optimal estimator is the $K=2$ estimator with the optimal order being $P=1$. When $N=150$, the optimal order was also $P=1$, but it was only marginally optimal relative to the $P=0$ order. Now, with an even larger sample size, the PDF of the optimal order is much tighter around its mode of $P=1$. Also, the frequency of $P=1$ has increased from .45 when $N=150$ to .65. Figures 2.19 and 2.20 show the results of the same model when $(\hat{\beta}_2, \hat{\beta}_3, \hat{\beta}_4, \hat{\beta}_5)$ has been doubled in magnitude. The optimal estimator is the $K=2$ estimator with the optimal order being $P=1$. Now, with X_1 and X_2 stronger in the outcome equation, the left tail of the PDF of the optimal order is essentially gone. Order values less than $P=1$ are virtually never selected. The right tail of the PDF is slightly fatter than before X_1 and X_2 were made stronger. Again, this is just as one should expect with a large sample size and strong covariates in the outcome equation.

Lastly, the results of the procedure on the dataset itself are examined. The known propensity score is again assumed to be $p = \frac{1}{N}\sum_{i=1}^{N} t_{Li} = .416$[23]. Table 2.1 presents the results of the GMM procedure. The estimates of the average treatment effect tend to decline with K. The $K = -1$ estimator and the $K = 0$ estimator are again identical since the true propensity score is assumed to equal the sample propensity score. The higher-order optimal order is $P = 1$. This corresponds to the $K = 2$ estimator. Thus, the standard difference-in-averages estimator is sub-optimal and an estimator similar to the HIR non-parametric propensity score estimator is also sub-optimal. There is a 10.5% difference in the higher-order optimal estimator relative to the more common difference-in-averages, $K = 0$, estimator. There is a 2.4% difference between the higher-order optimal estimator and the $K = 9$ estimator. This is a non-trivial difference, especially between the higher-order optimal estimator and the more common difference-in-averages estimator. Thus, implementation of the procedure of this section could likely have an important impact if the two-covariate specification was chosen.

The results of this section show that the procedure performs well in a more realistic setting in which an econometrician has data on a vector of covariates. Moment selection based on the $\hat{S}(K)$ criterion with a vector X performs well for the models considered and it appears to perform well for the actual experimental LaLonde dataset. It

[23] The results are very similar when the true propensity score is assumed to equal .5.

is interesting to note that the right tail of the optimal order PDF does not become as fat with sample size in the LaLonde Monte Carlo simulations of this section as it does in the standard Monte Carlo simulations of Section 2.3. In Section 2.3, a sample size of $N = 150$ could be considered large such that the penalty for including many moments in the estimation procedure is small. Thus, the PDF of the optimal order has an extremely fat tail. In this section, with a more realistic data generating process, even when $N = 445$ the right tails of the PDF graphs remain relatively trim. This seems to suggest that this procedure has merit over a non-trivial range of sample sizes in real world problems.

2.5 Conclusions

This chapter adjusts the framework of Chapter 1 to develop a higher-order optimal estimator for average treatment effects with experimental data and a vector of covariates. The key assumption of the chapter is that the econometrician pre-selects the subset of covariates to use in the estimation procedure. Given this selection, the higher-order optimal *polynomial order* of X is selected for inclusion as moments in the model. Theorem 2.1 shows that the results from the previous chapter are applicable in the case of a vector X.

The finite sample performance of the selection rule was again tested on simulated data from an artificial data generating process and on simulated data from a data generating process calibrated to LaLonde's 1986 dataset. The results again reinforced all previously developed intuition that large efficiency gains from this procedure are possible when sample size is small or when X is strong in the outcome equation. Finally, the

LaLonde calibrated models show that the selection rule has positive economic value over a significant range of sample sizes. In other words, a non-trivial range of sample sizes can be considered "finite" or "small," upon which the results of this chapter apply.

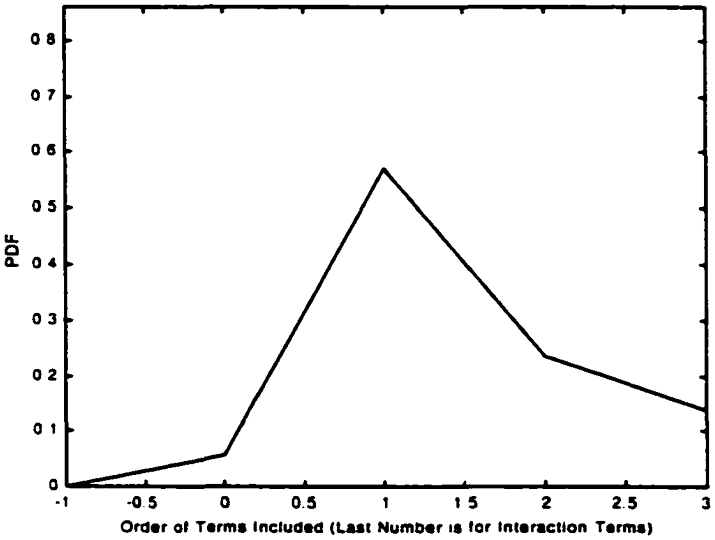

Figure 2.1: PDF of $\hat{P}$ Over 1000 Simulations (3 is interaction term)

($y = .5 + 2t + x_1 + x_2 + .4tx_1 + .6tx_2 + \varepsilon$, p(x) = 1/2, N = 25)

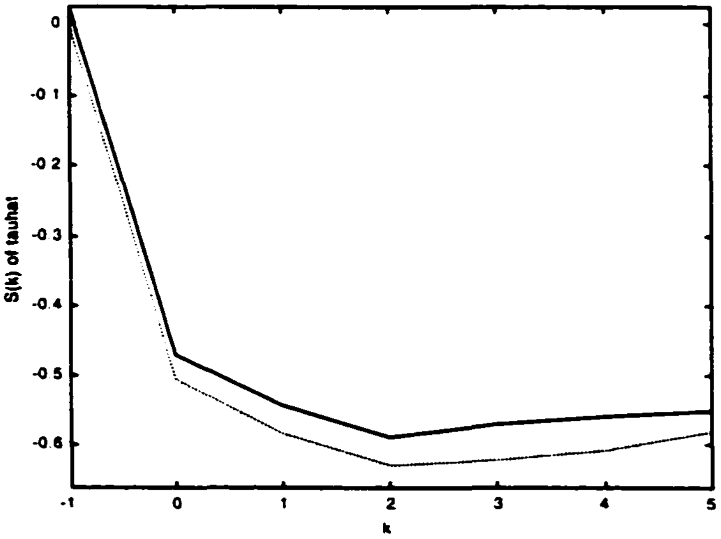

Figure 2.2: True $S(K)$ (Solid) and Mean of $\hat{S}(K)$ (Dotted) Over 1000 Simulations

($y = .5 + 2t + x_1 + x_2 + .4tx_1 + .6tx_2 + \varepsilon$, p(x) = 1/2, N = 25)

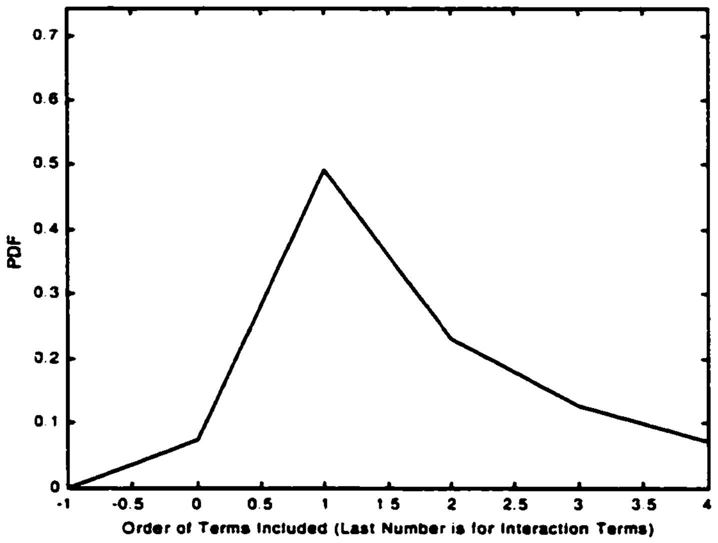

Figure 2.3: PDF of $\hat{P}$ Over 1000 Simulations (4 is interaction term)

($y = .5 + 2t + x_1 + x_2^2 + .4tx_1 + .6tx_2 + \varepsilon$, p(x) = 1/2, N = 25)

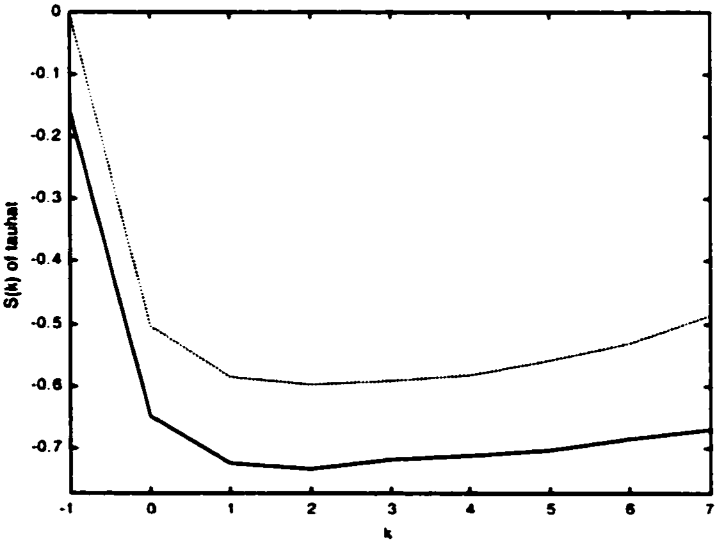

Figure 2.4: True $S(K)$ (Solid) and Mean of $\hat{S}(K)$ (Dotted) Over 1000 Simulations

($y = .5 + 2t + x_1 + x_2^2 + .4tx_1 + .6tx_2 + \varepsilon$, p(x) = 1/2, N = 25)

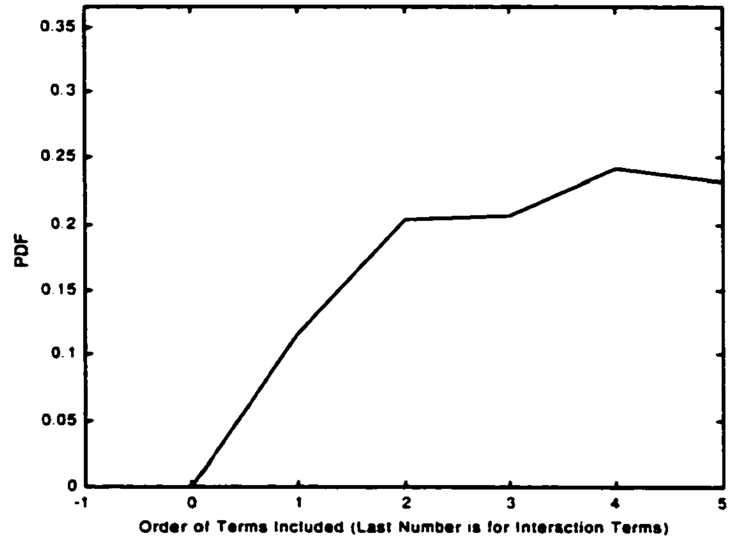

Figure 2.5: PDF of $\hat{P}$ Over 1000 Simulations (5 is interaction term)

($y = .5 + 2t + x_1 + x_2^2 + .4tx_1 + .6tx_2 + \varepsilon$, p(x) = 1/2, N = 150)

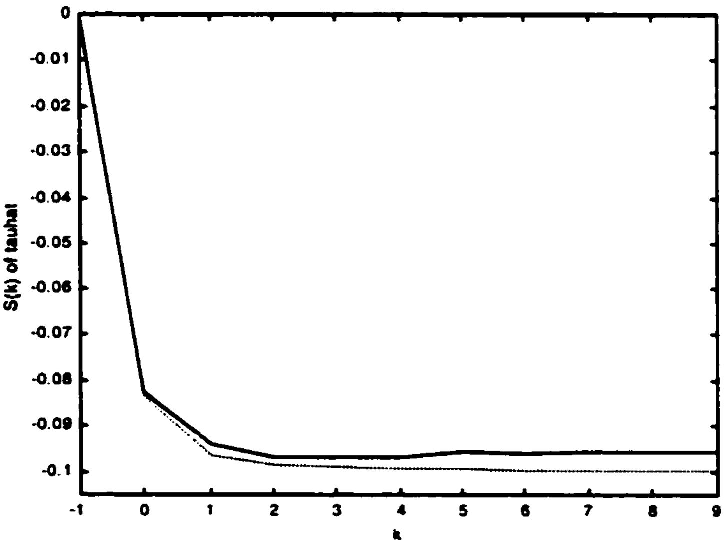

Figure 2.6: True $S(K)$ (Solid) and Mean of $\hat{S}(K)$ (Dotted) Over 1000 Simulations

($y = .5 + 2t + x_1 + x_2^2 + .4tx_1 + .6tx_2 + \varepsilon$, p(x) = 1/2, N = 150)

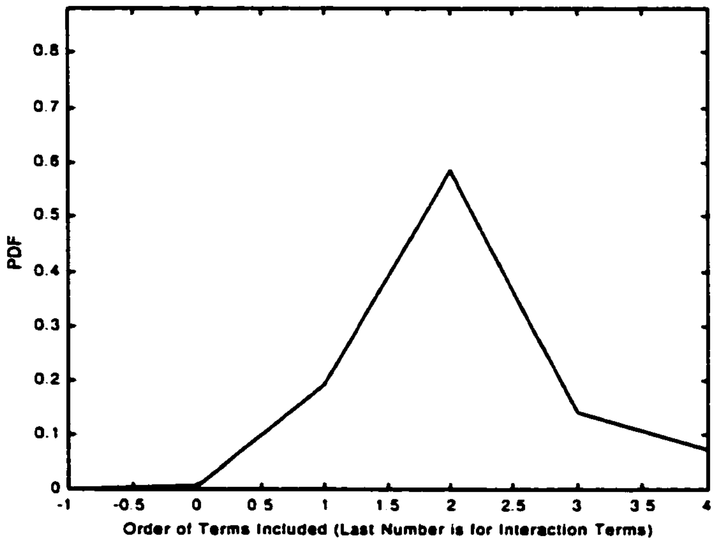

Figure 2.7: PDF of $\hat{P}$ Over 1000 Simulations (4 is interaction term)

$(y=.5+2t+x_1+x_2+7x_2^2+.4tx_1+.6tx_2+\varepsilon$, p(x) = 1/2, N = 25)

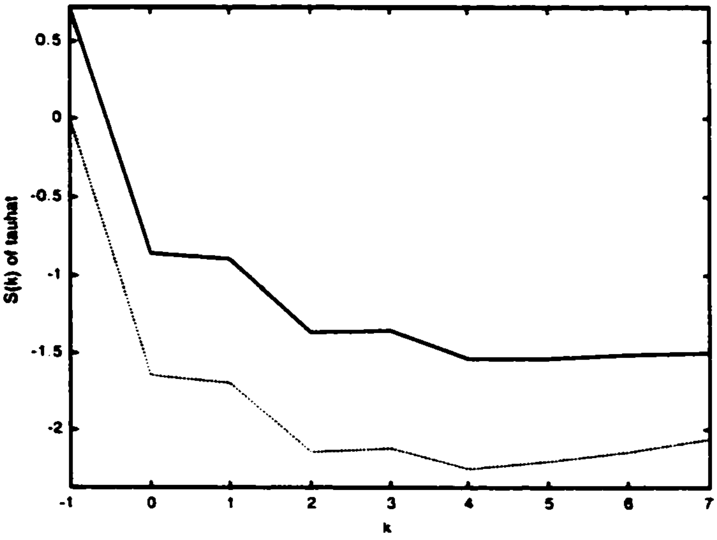

Figure 2.8: True $S(K)$ (Solid) and Mean of $\hat{S}(K)$ (Dotted) Over 1000 Simulations

$(y=.5+2t+x_1+x_2+7x_2^2+.4tx_1+.6tx_2+\varepsilon$, p(x) = 1/2, N = 25)

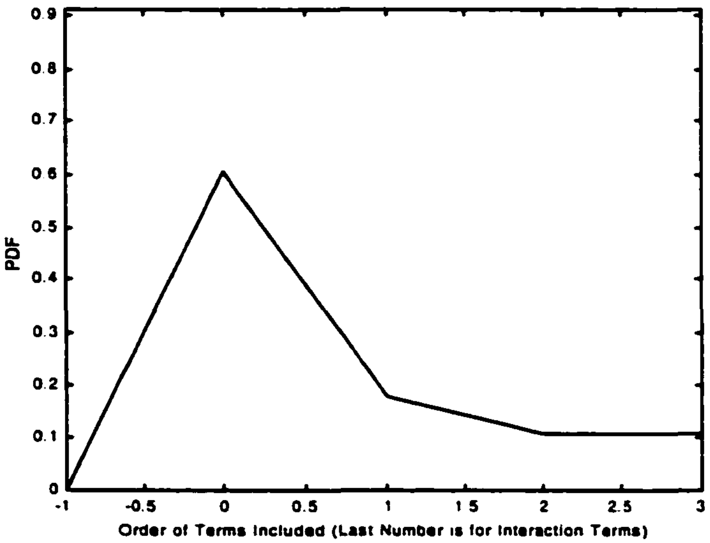

Figure 2.9: PDF of $\hat{P}$ Over 1000 Simulations (3 is interaction term)

($y = 3728 - 4772t$ $+.154x_1$ $+62.7x_2$ $+.027tx_1$ $+625tx_2$ $+\varepsilon$, p(x) = .416, N = 25)

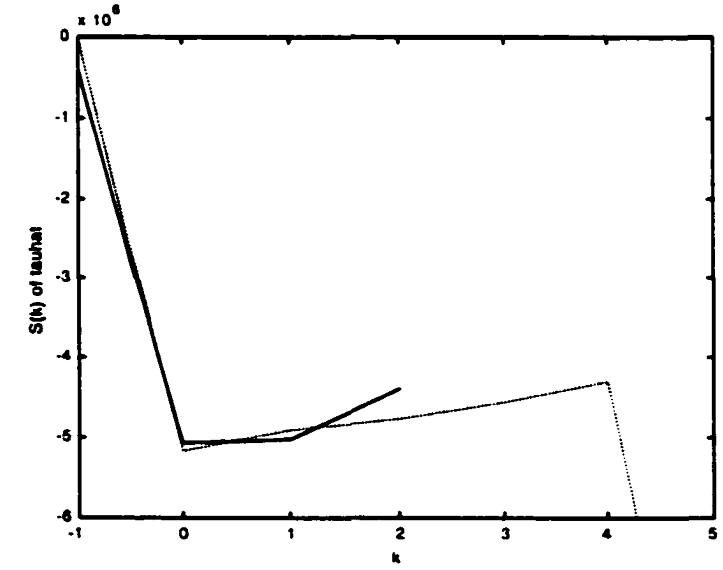

Figure 2.10: True $S(K)$ (Solid) and Mean of $\hat{S}(K)$ (Dotted) Over 1000 Simulations

($y = 3728 - 4772t$ $+.154x_1$ $+62.7x_2$ $+.027tx_1$ $+625tx_2$ $+\varepsilon$, p(x) = .416, N = 25)

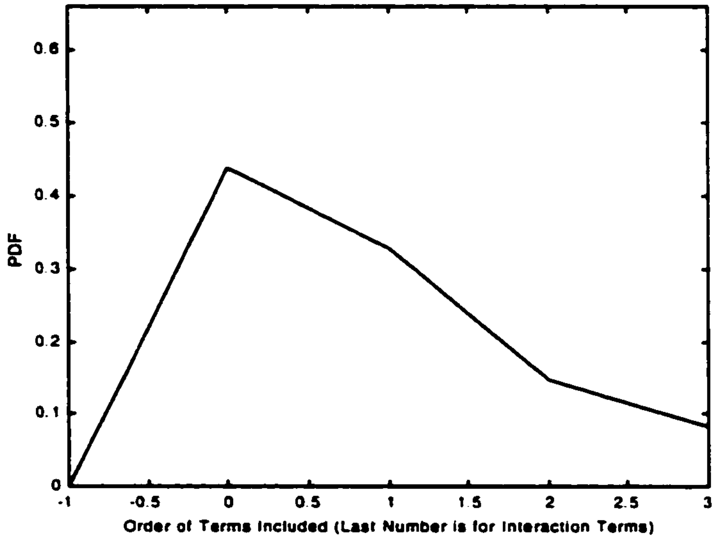

Figure 2.11: PDF of $\hat{P}$ Over 1000 Simulations (3 is interaction term)

($y = 3728 - 4772t + .307x_1 + 125x_2 + .054tx_1 + 1250tx_2 + \varepsilon$, p(x) = .416, N = 25)

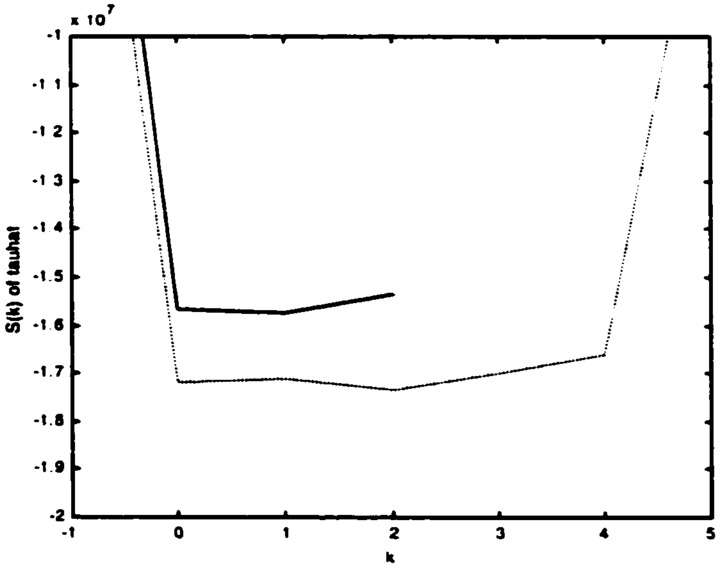

Figure 2.12: True $S(K)$ (Solid) and Mean of $\hat{S}(K)$ (Dotted) Over 1000 Simulations

($y = 3728 - 4772t + .307x_1 + 125x_2 + .054tx_1 + 1250tx_2 + \varepsilon$, p(x) = .416, N = 25)

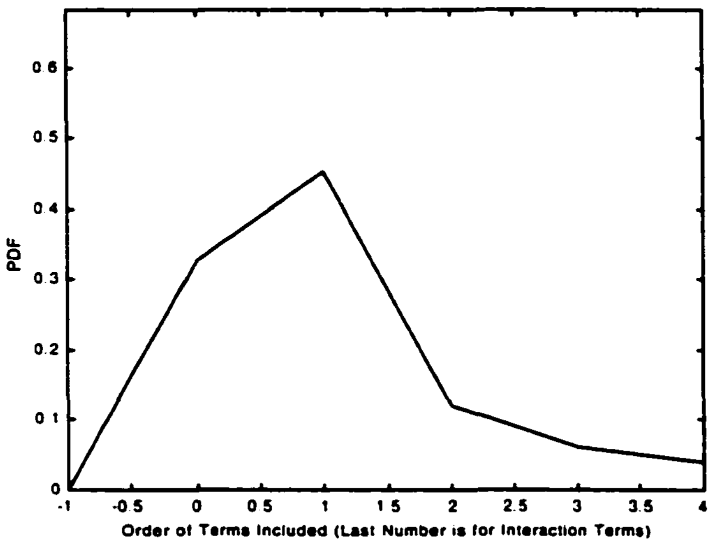

Figure 2.13: PDF of $\hat{P}$ Over 1000 Simulations (4 is interaction term)

($y = 3728 - 4772t + .154x_1 + 62.7x_2 + .027tx_1 + 625tx_2 + \varepsilon$, **p(x) = .416, N = 150)**

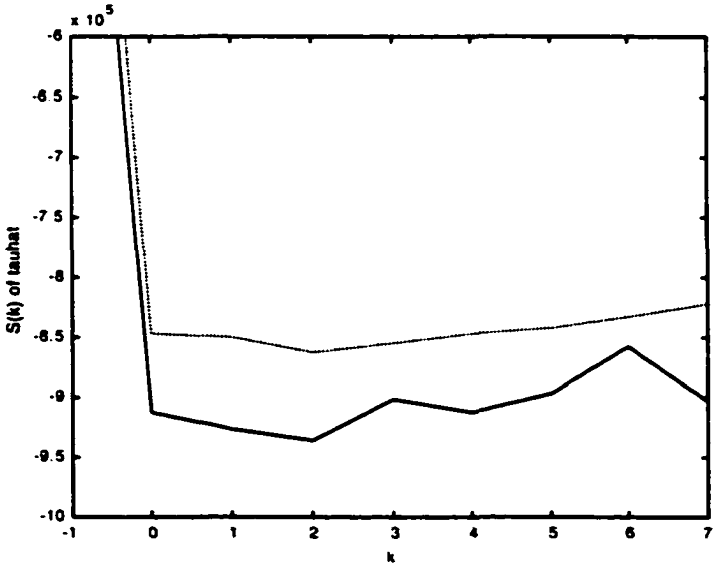

Figure 2.14: True $S(K)$ (Solid) and Mean of $\hat{S}(K)$ (Dotted) Over 1000 Simulations

($y = 3728 - 4772t + .154x_1 + 62.7x_2 + .027tx_1 + 625tx_2 + \varepsilon$, **p(x) = .416, N = 150)**

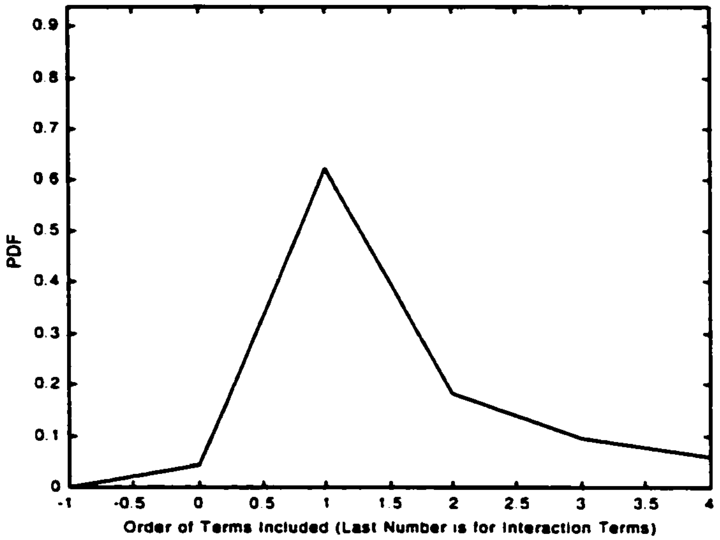

Figure 2.15: PDF of $\hat{P}$ Over 1000 Simulations (4 is interaction term)

($y = 3728 - 4772t + .307x_1 + 125x_2 + .054tx_1 + 1250tx_2 + \varepsilon$, p(x) = .416, N = 150)

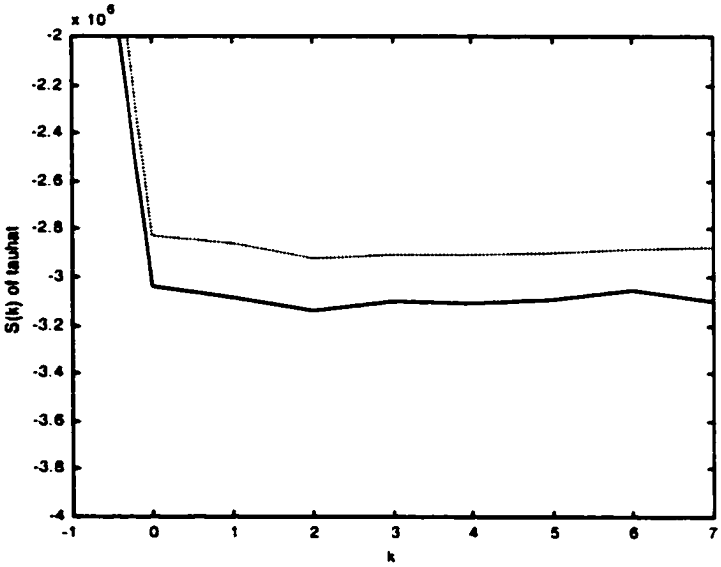

Figure 2.16: True $S(K)$ (Solid) and Mean of $\hat{S}(K)$ (Dotted) Over 1000 Simulations

($y = 3728 - 4772t + .307x_1 + 125x_2 + .054tx_1 + 1250tx_2 + \varepsilon$, p(x) = .416, N = 150)

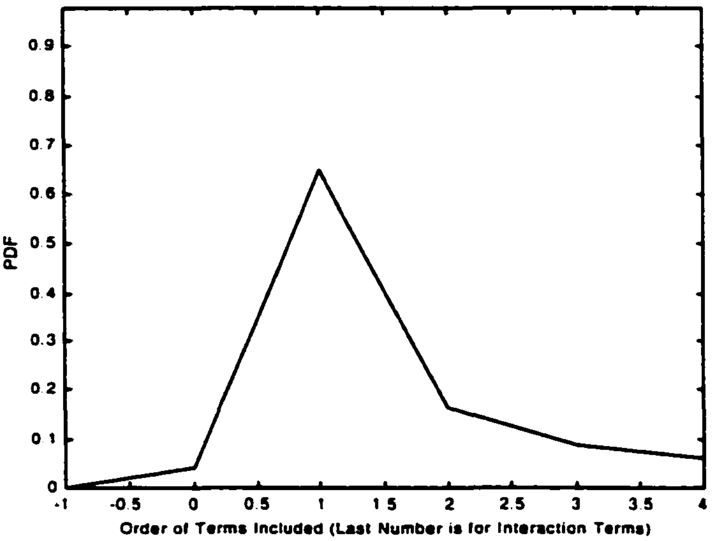

Figure 2.17: PDF of $\hat{P}$ Over 1000 Simulations (4 is interaction term)

($y = 3728 - 4772t + .154x_1 + 62.7x_2 + .027tx_1 + 625tx_2 + \varepsilon$, p(x) = .416, N = 445)

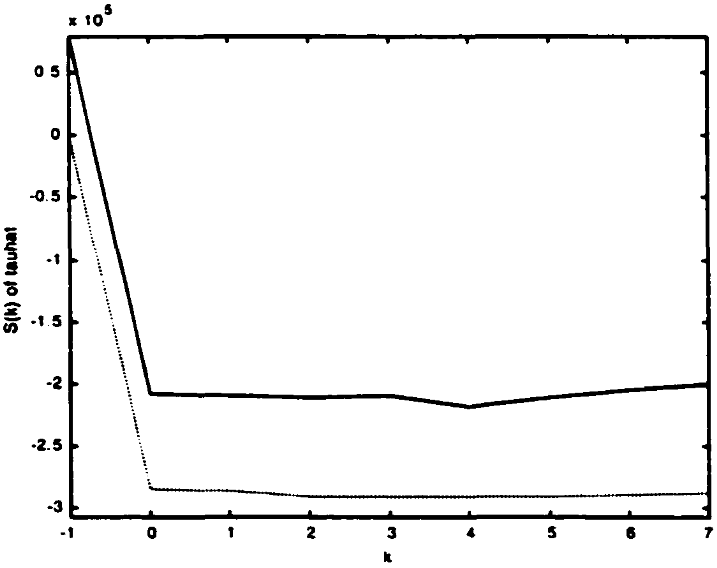

Figure 2.18: True $S(K)$ (Solid) and Mean of $\hat{S}(K)$ (Dotted) Over 1000 Simulations

($y = 3728 - 4772t + .154x_1 + 62.7x_2 + .027tx_1 + 625tx_2 + \varepsilon$, p(x) = .416, N = 445)

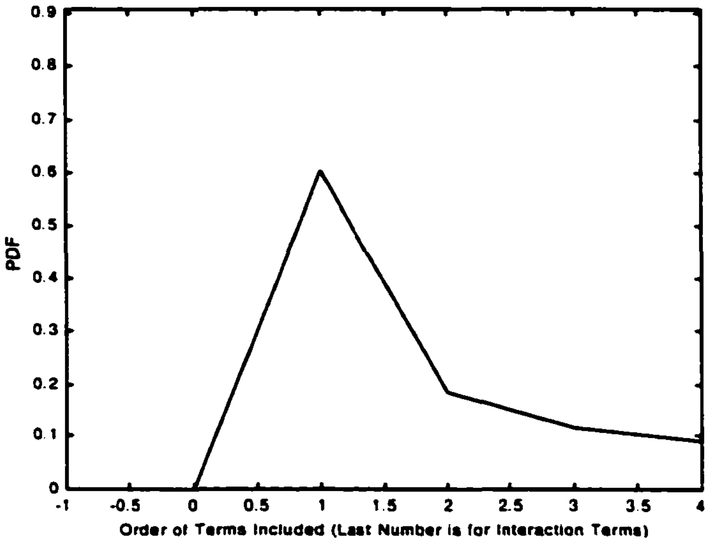

Figure 2.19: PDF of $\hat{P}$ Over 1000 Simulations (4 is interaction term)

($y = 3728 - 4772t + .307x_1 + 125x_2 + .054tx_1 + 1250tx_2 + \varepsilon$, p(x) = .416, N = 445)

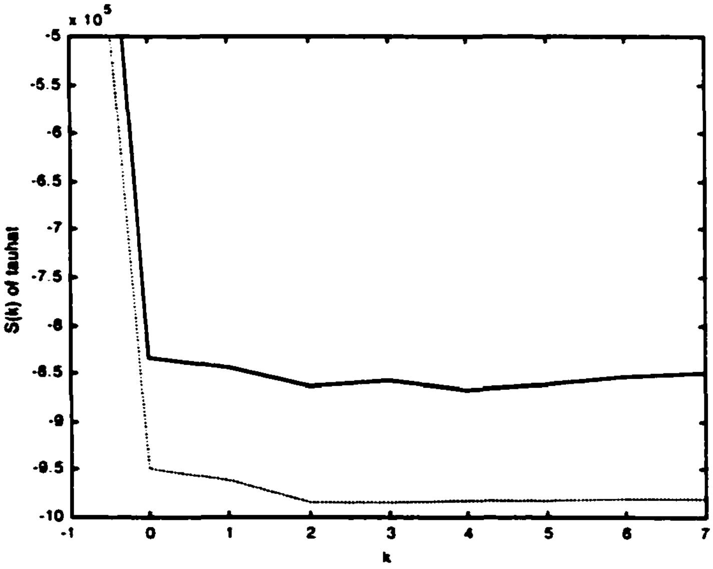

Figure 2.20: True $S(K)$ (Solid) and Mean of $\hat{S}(K)$ (Dotted) Over 1000 Simulations

($y = 3728 - 4772t + .307x_1 + 125x_2 + .054tx_1 + 1250tx_2 + \varepsilon$, p(x) = .416, N = 445)

K	Average Treatment Effect Estimate
-1	1794.3
0	1794.3
1	1748.9
2*	***1623.8***
3	1605.1
4	1406.8
5	1531.4
6	1536.1
7	1521.4
8	1502.9
9	1584.2

Table 2.1: Average Treatment Effect Estimates for a Sub-Sample of the NSW AFDC Participants of LaLonde (1986). Two covariates: Earnings in 1975 and Education.
***Optimal Order = 1 (Optimal K = 2).**

3. Chapter 3

Optimal Finite Sample Estimation of Binary Average Treatment Effects with Non-Experimental Data

3.1 Introduction

Analysis of optimal estimation of average treatment effects with non-experimental data is a valuable task. Many times, non-experimental data is all that is available to the econometrician. Not all job-training programs are offered experimentally through random selection. Certainly, a takeover of one firm over another is a function of various economic forces and factors, not a coin flip. Chapters 1 and 2 discuss the properties of a higher-order optimal GMM estimator of average treatment effects when the propensity score is known. However, in many, if not all, non-experimental settings the propensity score is not known. In this case, one observes the treatment status, pre-treatment variables, and outcome of all agents involved, but one does not know the form of the self-selection mechanism, the propensity score.

This chapter uses the analysis of the previous chapters to develop an optimal[24] GMM estimator of average treatment effects when the propensity score is unknown. The

[24] "Optimal" as opposed to "higher-order optimal" because no higher-order approximations are used.

previous GMM moments are adjusted to account for the fact that the propensity score is know longer known and must be estimated in some manner. Logit estimation of the propensity score is proposed where the terms in the Logit regression _and_ the GMM moments are selected simultaneously. In this setting, the question of interest is: What is the optimal number of terms to include in the Logit regression and the GMM moments so that the MSE of the resulting estimator will be minimized? Just like in previous chapters, the term "optimal" is relative to all possible estimators within the given GMM class. The complicated form of the MSE of the resulting estimator will be shown, and then a simple approximation to this MSE will be derived. The approximation is not an asymptotically optimal one (as in the previous chapters) although it is a simple one. Finally, using artificial Monte Carlo simulations and simulations calibrated to the non-experimental data of LaLonde (1986), the finite sample MSE of the newly derived Unknown Propensity Score GMM estimator will be compared to two forms of the HIR non-parametric propensity score estimator. One form uses cross-validation for bandwidth selection and the other uses Silverman's Rule of Thumb.

It will be shown that in the majority of cases the optimal Unknown Propensity Score GMM Estimator outperforms the HIR estimator with respect to finite sample MSE. This efficiency gain converges to zero as sample size increases. However, the selection criterion for choosing the optimal K is shown to be unreliable. Also, the finite sample performance of the latter two estimators will be compared to the performance of the optimal Known Propensity Score GMM Estimator of Chapters 1 and 2. The Known Propensity Score GMM estimator outperforms the two unknown propensity score estimators in all cases, and the efficiency gain also converges to zero as sample size increases. Hence, relative to the two unknown propensity score estimators considered, the propensity score is not ancillary for estimation of average treatment effects in finite samples[25]. This result is the finite sample compliment to the asymptotic results of Hahn (1998), which state that the propensity score is ancillary for estimation of average treatment effects.

The format of this chapter is as follows. Section 3.2 sets up the adjusted GMM framework, derives the Unknown Propensity Score GMM estimator, and develops the moment selection criterion to be used. Section 3.3 compares the performance of the three estimators discussed when X is a scalar. This is done with artificial Monte Carlo simulations and non-experimental LaLonde data calibrated simulations. Section 3.4 presents results of similar comparisons to Section 3.3, except X is a 2×1 vector. Finally, Section 3.5 concludes.

3.2 Adjusted GMM Framework

This section presents the details of the adjusted GMM framework proposed for use in estimation of average treatment effects when the propensity score is unknown. Given that the econometrician has no knowledge of the propensity score, the estimation method of Chapters 1 and 2 is no longer feasible. Thus, the adjusted GMM moment conditions become

[25] See Hahn (1998) for asymptotic results.

$$\tilde{\psi}(y,t,x,\tau)=\begin{pmatrix}\tilde{\psi}_1(y,t,x,\tau)\\ \tilde{\psi}_2(t,x)\end{pmatrix},$$

where

$$\tilde{\psi}_1(y,t,x,\tau)=\left(\frac{y\cdot t}{p(x)}-\frac{y\cdot(1-t)}{1-p(x)}\right)-\tau=V-\tau,$$

$$\tilde{\psi}_2(t,x)=\begin{pmatrix}t-p(x)\\ f_1(x)(t-p(x))\\ \vdots\\ f_P(x)(t-p(x))\end{pmatrix},$$

$$E(\tilde{\psi}(y,t,x,\tau))=0,$$

and $p(x)$ is no longer known. Let $K+1$ be the total number of moments included in $\tilde{\psi}_2$. For the remainder of the chapter X will be added to the auxiliary moments in ascending polynomial order, whether X is a scalar or a vector. Covariate interactions were discussed in Chapter 2. They are not treated in this chapter because the main focus is to compare and contrast the finite sample performance of three estimators: the optimal Known Propensity Score GMM estimator, the optimal Unknown Propensity Score GMM estimator, and the HIR estimator. Thus, let P be the total number of polynomial orders of X included in $\tilde{\psi}_2$. When X is a scalar, $K=P$, and when X is a $D\times 1$ vector, $K=D\cdot P$.

If the propensity score is known, the setup above is identical to that of Chapter 1 when X is a scalar and Chapter 2 when X is a vector. Since the propensity score is unknown, the following identification assumption is made. Let

$$p(x)=\frac{e^{\alpha_0+\alpha_1' f_1(x)+\cdots+\alpha_P' f_P(x)}}{1+e^{\alpha_0+\alpha_1' f_1(x)+\cdots+\alpha_P' f_P(x)}}.$$

If X is a scalar then $f_i(x)=x^i$ and

$$p(x)=\frac{e^{\alpha_0+\alpha_1 x+\cdots+\alpha_K x^K}}{1+e^{\alpha_0+\alpha_1 x+\cdots+\alpha_K x^K}},$$

where α_i is an unknown scalar parameter to be estimated. If X is a vector then $f_i(x)=x^i$ and

$$p(x)=\frac{e^{\alpha_0+\alpha_1' x+\cdots+\alpha_P' x^P}}{1+e^{\alpha_0+\alpha_1' x+\cdots+\alpha_P' x^P}},$$

where x^i is $D\times 1$ and α_i is an unknown $D\times 1$ parameter vector to be estimated. Hence, the propensity score is parameterized by $K+1=D\cdot P+1$ unknown parameters. Plugging the propensity score into $\tilde{\psi}=\tilde{\psi}(y,t,x,\tau,\alpha)$ yields a system of $K+2$ moment conditions and $K+2$ unknowns. This system of moments is always just-identified. Intuitively, it is assumed that, since one does not have information on the propensity score or the optimal moments to include in the estimation procedure, one includes a marginal term of X in a moment restriction and in the Logit form of the propensity score simultaneously. Assuming a Logit functional form for the propensity score has two nice properties. First, a Logit with higher-order polynomials of X in the exponential can approximate any continuous PDF arbitrarily closely. Second, the fitted probabilities are always contained in the interval $(0,1)$. The Linear Probability Model also has the first property but does not have the second. Use of the Linear Probability Model will be discussed in the results below.

The estimation procedure for the adjusted GMM framework is as follows. $\tilde{\psi}$ is just-identified so there will be a unique solution for the estimators $(\tilde{\tau},\tilde{\alpha})$. Notice that when $p(x)$ is plugged in, $\tilde{\psi}_2$ is exactly equal to the Maximum Likelihood moments of the Logit[26] and is only a function of α. Thus, α is estimated by running the appropriate Logit regression. The resulting estimate of $\tilde{\alpha}$ is then plugged into $\tilde{\psi}_1$ to solve for $\tilde{\tau}$. Specifically,

$$\tilde{\tau}=\frac{1}{N}\sum_{i=1}^{N}\frac{y_i\cdot t_i}{\tilde{p}(x_i)}-\frac{y_i\cdot(1-t_i)}{1-\tilde{p}(x_i)}=\frac{1}{N}\sum_{i=1}^{N}\tilde{V}_i, \tag{23}$$

where $\tilde{p}(x)$ is the fitted Logit probability. If $\tilde{p}(x)$ is estimated non-parametrically, $\tilde{\tau}$ will be the HIR non-parametric propensity score estimator. The adjusted GMM framework of this chapter essentially substitutes a flexible Logit probability estimate in place of a non-parametric probability estimate. The HIR estimator includes either all or none of the information contained in X. In finite samples, there may be an efficiency gain to including a subset of this information, in the form of lower-order polynomials of X, in a flexible Logit specification.

The next step is to develop a criterion upon which to select K or P. Hence, an approximation to the MSE of the resulting estimator is necessary. With the HIR estimator this step is not necessary, either X is fully included in estimation of the

[26] The score function, whose expectation is zero.

propensity score or it is not. However, optimal bandwidth selection becomes an important issue. Remember, the main goal of this dissertation is to develop an estimator that minimizes the MSE of the average treatment effect estimator under a variety of assumptions (chapters). The adjusted GMM framework of this chapter yields the estimator of equation (23). Subtracting τ yields

$$\tilde{\tau}-\tau=\frac{1}{N}\sum_{i=1}^{N}\left(\tilde{V}_i-\tau\right).$$

Thus, the MSE of $\tilde{\tau}$ is

$$E\left(\left(\tilde{\tau}-\tau\right)^2\right)=E\left(\left(\frac{1}{N}\sum_{i=1}^{N}\left(\tilde{V}_i-\tau\right)\right)^2\right)=\frac{1}{N^2}E\left(\sum_{i=1}^{N}\sum_{j=1}^{N}\left(\tilde{V}_i-\tau\right)\left(\tilde{V}_j-\tau\right)\right)$$

$$=\frac{1}{N}Var\left(\tilde{V}_i\right)+\frac{1}{N^2}\sum_{i=1}^{N}\sum_{j\neq i}Cov\left(\tilde{V}_i,\tilde{V}_j\right) \qquad . \qquad (24)$$

If the $\tilde{V}_i$'s are independent, the MSE of $\tilde{\tau}$ reduces to $\frac{Var\left(\tilde{V}_i\right)}{N}$ where $Var\left(\tilde{V}_i\right)$ can be approximated with the sample variance of $\tilde{V}_i$. This would provide an asymptotically optimal approximation to the true MSE. However, the $\tilde{V}_i$'s are not independent because each is a function of the estimator $\tilde{\alpha}$. Thus, the true MSE of $\tilde{\tau}$ is a function of $Var\left(\tilde{V}_i\right)$ as well as $Cov\left(\tilde{V}_i,\tilde{V}_j\right)$ for all $j>i$. Estimating all relevant covariances in practice can be very difficult so the formula for the MSE of $\tilde{\tau}$ when the $\tilde{V}_i$'s are independent will be used throughout the remainder of the chapter for selection of K. Specifically, let

$$\widetilde{MSE}(K)=\tilde{S}(K)=\frac{Var\left(\tilde{V}_i\right)}{N}=\frac{1}{N^2}\sum_{i=1}^{N}\left(\tilde{V}_i(K)-\bar{V}(K)\right)^2, \qquad (25)$$

where $\bar{V}(K)=\frac{1}{N}\sum_{i=1}^{N}\tilde{V}_i(K)$. Let $\tilde{K}=\underset{K}{\operatorname{argmin}}\{\tilde{S}(K)\}$. $\tilde{S}(K)$ is not asymptotically optimal like $\hat{S}(K)$ in Chapters 1 and 2[27]. However, it has a benefit in that it is extremely simple to calculate and use in practice. The performance of the $\tilde{S}(K)$ moment selection criterion will be examined in the Monte Carlo simulations in the next two sections.

3.3 Monte Carlo Simulations with a Scalar Covariate

This section presents simulation results for the case where X is a scalar. Two types of simulation are performed, artificial Monte Carlo simulations and Monte Carlo simulations calibrated to a subset of LaLonde's 1986 non-experimental dataset on the effects of a job-training program on future earnings. The finite sample MSE of the Known Propensity Score GMM estimator of Chapters 1 and 2, the Unknown Propensity Score GMM estimator of this chapter, and the HIR estimator are presented and compared for all data generating process considered.

3.3.1 Artificial Monte Carlo Simulations

The following artificial data generating process is used for the simulations of this section. Let

$$Y_i = \beta_0 + \beta_1 T_i + \beta_2 X_i + \beta_3 T_i X_i + \varepsilon_i ,$$

$$\varepsilon_i \sim N(0,1),$$

[27] $\hat{S}(K)$ is higher-order asymptotically optimal. There are no higher-order MSE approximations in this case, so the term "higher-order" is left out.

$X_i \sim UNIF(-1,1)$,

$$p(X_i) = \frac{e^{\alpha_0+\alpha_1 X_i +\alpha_2 X_i^2 +\alpha_3 X_i^3}}{1+e^{\alpha_0+\alpha_1 X_i +\alpha_2 X_i^2 +\alpha_3 X_i^3}},$$

where $(\beta_0,\beta_1,\beta_2,\beta_3)=(.5,2,1,.4)$ and $(\alpha_0,\alpha_1,\alpha_2,\alpha_3)=(.1,.7,-.4,.3)$. In this data generating process, assignment to treatment is no longer random and the treatment effect varies across individuals through the interaction of T with X. The average treatment effect, in this case, is equal to $\tau = \beta_1 + \beta_3 E(X) = 2$. In what follows, the MSE of the Known Propensity Score GMM estimator and the Unknown Propensity Score GMM estimator is calculated as the sample MSE of 5,000 average treatment effect estimates for each K value. The MSE of the HIR non-parametric propensity score estimator is calculated as the sample MSE of 1,000 average treatment effect estimates for each estimation method (cross-validation and Silverman's Rule of Thumb)[28]. Also, the $\tilde{S}(K)$ criterion is used to select the optimal K value, $\tilde{K}$, for 1,000 samples. The PDF of these $\tilde{K}$ values is plotted to show the performance of the $\tilde{S}(K)$ moment selection criterion for the various data generating processes.

Table 3.1 and Figure 3.1 present results for the case where $N = 50$. The optimal Known Propensity Score GMM estimator is the $K = 1$ estimator. The optimal Unknown Propensity Score GMM estimator is also the $K = 1$ estimator. The optimal HIR estimator uses Silverman's Rule of Thumb for bandwidth selection. This is not surprising considering that Silverman's Rule of Thumb is the optimal bandwidth selection rule in one dimension. It's two dimensional performance will be examined in the next section. Figure 3.1 shows the PDF of $\tilde{K}$ for the Unknown Propensity Score GMM estimator. It can be seen that the $\tilde{S}(K)$ criterion performs relatively poorly although the mode of the PDF is at the true optimal K value, $K = 1$. Table 3.1 shows that, in the current data generating process, the optimal Known Propensity Score GMM estimator has the lowest MSE, followed by the optimal Unknown Propensity Score GMM estimator, and then the optimal HIR estimator. The optimal Known Propensity Score GMM estimator yields a 3.9% MSE reduction relative to the optimal Unknown Propensity Score GMM estimator and a 6.0% MSE reduction relative to the optimal HIR estimator. Thus, knowledge of the propensity score is non-trivial relative to one of the more common estimators used when the propensity score is unknown, the HIR estimator. Also, the Unknown Propensity Score GMM estimator developed earlier in this chapter has a 2.3% MSE reduction relative to its unknown propensity score counterpart, the HIR estimator. Hence, in this case, because of the small sample size, if the propensity score is unknown, it is optimal to include only a subset of the information contained in X in the estimation procedure rather than all of it.

Table 3.2 and Figure 3.2 present results for the case where sample size is increased to $N = 150$. The optimal Known Propensity Score GMM estimator increases to the $K = 3$ estimator, as expected. The optimal Unknown Propensity Score GMM estimator remains the $K = 1$ estimator. Finally, the optimal HIR estimator still uses

[28] 1,000 iterations, rather than 5,000 iterations, was chosen because cross-validation routines can take a long time to run in practice if the potential bandwidth vector is significantly exhaustive.

Silverman's Rule of Thumb for bandwidth selection. Figure 3.2 shows that the PDF of $\tilde{K}$ for the Unknown Propensity Score GMM estimator becomes tighter around the true optimal K value, $K = 1$. However, the distribution is still relatively thick. Table 3.2 shows that, again, the optimal Known Propensity Score GMM estimator has the lowest MSE, followed by the optimal Unknown Propensity Score GMM estimator, and then the optimal HIR estimator. Now, the optimal Known Propensity Score GMM estimator yields a smaller 3.6% MSE reduction relative to the optimal Unknown Propensity Score estimator and a smaller 4.6% MSE reduction relative to the optimal HIR estimator. Also, as expected, the MSE of the optimal Unknown Propensity Score GMM estimator and the optimal HIR estimator become closer in percentage terms (there is a 1.1% MSE difference). Thus, there is still a finite sample advantage for the optimal Known Propensity Score GMM estimator over the two unknown propensity score estimators, but the advantage has diminished. Also, the advantage of the optimal Unknown Propensity Score GMM estimator over the optimal HIR estimator has diminished.

Table 3.3 and Figure 3.3 present results for the case where sample size is increased even more to $N = 500$. In this case, the optimal Known Propensity Score GMM estimator remains the $K = 3$ estimator and the optimal Unknown Propensity Score GMM estimator increases to the $K = 2$ estimator. The optimal HIR estimator still uses Silverman's Rule of Thumb. Figure 3.3 shows that the PDF of $\tilde{K}$ for the Unknown Propensity Score GMM estimator now becomes significantly tight around the true optimal K value, $K = 2$. From Table 3.3, it can be seen that the optimal Known Propensity Score GMM estimator and the optimal Unknown Propensity Score GMM estimator perform very similar. There is only a .6% difference in their MSEs. The difference in performance between the optimal Known Propensity Score GMM estimator and the optimal HIR estimator diminishes to 2.5%, and the difference in performance between the optimal Unknown Propensity Score GMM estimator and the optimal HIR estimator is small at 1.9%.

Asymptotically, the MSEs of the three estimators should tend to converge to the semi-parametric efficiency bound[29]. The results indicate that the MSEs are converging as sample size increases. However, in the finite sample data generating processes above, the optimal Known Propensity Score GMM estimator always outperforms the two unknown propensity score estimator and the optimal Unknown Propensity Score GMM estimator always outperforms the optimal HIR estimator. The largest MSE reduction is 6%. While this is non-trivial, one may expect this number to be even larger in truly small samples or samples with equivalent sample sizes and a vector of covariates. Remember from Chapters 1 and 2 that "large sample" results tended to obtain relatively quickly with the artificial Monte Carlo data. This, combined with the use of a scalar covariate, is why relatively small percentage MSE decreases are observed in the setting above in what would initially appear to be a "small sample size". Also, the $\bar{S}(K)$ criterion for choosing the optimal K value for the Unknown Propensity Score GMM estimator performs poorly when sample size is 50 and 150, and much better when sample size is 500. Thus, while the optimal Unknown Propensity Score GMM estimator has its most significant improvement over the HIR estimator in small sample sizes, the selection criterion for choosing the optimal K value is poor over these ranges. This is due to the fact that $\bar{S}(K)$ leaves out relevant covariances in its functional form. In small samples, these ignored lower-order covariance terms can be significant.

3.3.2 Monte Carlo Simulations Calibrated to LaLonde Data

This section performs similar experiments as Section 3.3.1 except the Monte Carlo fake data generation mechanism is calibrated to a subset of the non-experimental dataset of LaLonde (1986). In this dataset, the treated individuals remain the treated NSW AFDC participants. However, the controls are now drawn from the Panel Study of Income Dynamics (PSID). Thus, the dataset exemplifies a non-experimental sample. Table 3.4 shows the sample means and standard deviations for the outcome variable (Earnings in 1978), the treatment variable, and all pre-treatment variables of this dataset. Means and standard deviations are also shown separately for the treated and for the controls. In this sample, selection into treatment status is not random. The sample means and standard deviations of the pre-treatment variables (Earnings in 1975 and below) reflect this fact in that they are not similar. The two closest means occur in the Age and Education variables. This is one of the reasons why these variables are chosen for use in the LaLonde simulation routine. The other reasons are discussed below.

A Monte Carlo simulation is calibrated to the LaLonde dataset according to the following rules. They are similar to the rules of Chapter 1. Let x_L be the observed scalar covariate. Let (y_L, t_L) be the observed outcome vector and treatment vector from

[29] When normalized.

LaLonde's dataset, respectively. Run a Logit regression of t_L on $\left(1, x_L, x_L^2\right)$ to get estimates of the resulting parameter vector $\hat{\alpha} = \left(\hat{\alpha}_0, \hat{\alpha}_1, \hat{\alpha}_2\right)'$. The fitted Logit probabilities will be the true propensity score values for each observed x_L. Run the OLS regression of

$$y_L = \beta_0 + \beta_1 t_L + \beta_2 x_L + \beta_3 t_L x_L + \varepsilon .$$

Let $\hat{\beta}$ be the resulting vector of parameter estimates and $\hat{\sigma}^2 = \hat{\varepsilon}'\hat{\varepsilon} / (N-4)$. Draw x from the empirical distribution of x_L. Calculate the true propensity score, $p(x)$, as the fitted Logit probability of $t \mid x$. Draw t from a $BIN\left(1, p(x)\right)$ distribution. Draw v from a $N\left(0, \hat{\sigma}^2\right)$ distribution. Finally, let

$$y = \hat{\beta}_0 + \hat{\beta}_1 t + \hat{\beta}_2 x + \hat{\beta}_3 t x + v .$$

A key assumption in the estimation of average treatment effects is Assumption 4, the assumption that the propensity score is not close to zero or one. This implies that the data contains individuals with identical or similar covariates, some who received the treatment and some who did not. Assumption 1 (Unconfoundedness) then implies that these individuals are comparable with respect to their potential outcomes. An implication of the propensity score being bounded away from zero and one is that a histogram plot of X for the treated should have significant overlap with a histogram plot of X for the controls.

Figure 3.4 shows these histogram plots for the three non-indicator covariates Earnings in 1975, Education, and Age. One can see that there is essentially no overlap in the distribution of Earnings in 1975 between the treated and the controls. There is small but significant overlap in the distribution of Education in the 8 to 12 range. Finally, there is again small but significant overlap in the distribution of Age in the 18 to 35 range. The covariate with the best overlap is Age, and this is the reason why Age is chosen as the scalar covariate of this section. Even though Age has the best overlap in its distribution, the true propensity score (as derived above) was still very often close to zero. Its minimum value was .011, its maximum value was .336, and its mean was .069. In theory this may be considered bounded from zero, but in practice, with small sample sizes, this results in many singular matrices and highly unreliable and volatile results. Because of this the following changes were made. Observations with an Age value less than 18 or greater than 33 were deleted from the LaLonde dataset. This resulted in a decrease in sample size from 2675 to 1436. Even with this adjustment, the propensity score remained low so the $\hat{\alpha}_1$ parameter was decreased to 90% of its original value. The new true propensity score has minimum value of .428, a maximum value of .684, and a mean of .482. Thus, the distribution of the true propensity score has been centered and significantly (for practical small sample purposes) bounded away from zero and one. Figure 3.5 shows the histogram of the original pre-modified propensity score and the histogram of the post-modified propensity score.

Table 3.5 and Figure 3.6 present results for the LaLonde case where $N = 50$. The optimal Known Propensity Score GMM estimator is the $K = 1$ estimator. The optimal Unknown Propensity Score GMM estimator is the $K = 0$ estimator and the optimal HIR estimator uses cross-validation for bandwidth selection. Figure 3.6 shows that the PDF of $\tilde{K}$ has its mode at the optimal value, $K = 0$, but is thick-tailed to the right. From Table 3.5, one can see that the optimal Known Propensity Score GMM estimator has the lowest MSE. This time it is followed by the optimal HIR estimator and then the optimal Unknown Propensity Score GMM estimator. The $K = 0$ estimator in the Unknown Propensity Score GMM estimator of this chapter uses a Logit regression of the treatment indicator on a constant. The result is identical to an OLS regression on a constant in that the fitted probabilities are equal to the sample mean of the treatment indicator, $\hat{t} = \frac{1}{N}\sum_{i=1}^{N} t_i$. These fitted probabilities are then plugged into the weighting equation, $\tilde{V}$, to obtain the average treatment effect estimator. Hence, the $K = 0$ estimator is identical to the simple difference-in-averages estimator, $\hat{\tau}_s$, of Chapter 1. This estimator can also be obtained through non-parametric estimation of the propensity score with an infinitely large bandwidth[30]. However, a cross-validation routine selects the optimal bandwidth over a range of possible values. If one of these potential bandwidth values is large, the cross-validation routine should never perform worse than the $K = 0$ Unknown Propensity Score GMM estimator. This is because the $K = 0$ Unknown Propensity Score GMM estimator is included in the set of the possible non-parametric estimators from which the cross-validation routine chooses! Hence, in Table 3.5, the HIR estimator from cross-validation has a lower MSE than the optimal $K = 0$ Unknown Propensity Score GMM estimator. However, knowledge of the propensity score would still improve efficiency.

[30] Or a finite but large bandwidth in practice.

Table 3.6 and Figure 3.7 show the results for the case where sample size is increased to $N = 150$. The optimal Known Propensity Score GMM estimator remains the $K = 1$ estimator. The optimal Unknown Propensity Score GMM estimator increases to the $K = 1$ estimator. Finally, the optimal HIR estimator now uses Silverman's Rule of Thumb. From Figure 3.7, it can be seen that the right tail of the PDF of $\tilde{K}$ is now thinner. However, the mode of the PDF is now at the sub-optimal $K = 0$ value. This illustrates the unreliability of the $\tilde{S}(K)$ criterion in this case. Table 3.6 shows that the optimal Known Propensity Score GMM estimator has the lowest MSE, followed by the optimal Unknown Propensity Score GMM estimator, and then the optimal HIR estimator. The optimal Known Propensity Score GMM estimator yields a 3.0% MSE reduction relative to the optimal Unknown Propensity Score GMM estimator and a 4.2% MSE reduction relative to the optimal HIR estimator. Also, the MSE of the optimal HIR estimator is now greater than the MSE of the optimal Unknown Propensity Score GMM estimator (the latter contains a 1.2% reduction). This is because sample size has increased, resulting in the $K = 0$ estimator no longer being the optimal Unknown Propensity Score GMM estimator[31].

Table 3.7 and Figure 3.8 present results for the case where sample size is increased even more to $N = 500$. Again, the performance ordering of the three estimators remains the same. The MSE of the three estimators is converging as sample size increases. Finally, Figure 3.3 shows that the PDF of $\tilde{K}$ for the Unknown Propensity

[31] Note that the MSE of both non-parametric estimators is less than the $K = 0$ Unknown Propensity Score GMM estimator.

Score GMM estimator becomes significantly tight around the true optimal K value, $K = 1$. Thus, this section reinforces the intuition developed in Section 3.3.1. The only deviation occurs when sample size is low and X is sufficiently weak such that the optimal Unknown Propensity Score GMM estimator is at $K = 0$. In this situation, it is necessarily the case that the HIR estimator with cross-validation will outperform the Unknown Propensity Score GMM estimator.

3.4 Monte Carlo Simulations with a Vector of Covariates

This section presents simulation results for the case where X is a 2×1 vector. The optimal polynomial order of X, P, is now the relevant choice variable in the two GMM estimation procedures Again, an artificial Monte Carlo simulation and a Monte Carlo simulation calibrated to a subset of LaLonde's 1986 non-experimental dataset are performed and the results are discussed.

3.4.1 Artificial Monte Carlo Simulations

The data generating process of Section 3.3.1 is slightly adjusted to allow for the second covariate. Now, let

$$Y_i = \beta_0 + \beta_1 T_i + \beta_2 X_{1i} + \beta_3 X_{2i} + \beta_4 T_i X_{1i} + \beta_5 T_i X_{2i} + \varepsilon_i ,$$

$$\varepsilon_i \sim N(0,1),$$

$$X_{1i} \sim UNIF(-1,1), X_{2i} \sim UNIF(-.5,.9),$$

$$p(X_i) = \frac{e^{\alpha_0 + \alpha_1 X_{1i} + \alpha_2 X_{2i} + \alpha_3 X_{1i}^2 + \alpha_4 X_{2i}^2}}{1 + e^{\alpha_0 + \alpha_1 X_{1i} + \alpha_2 X_{2i} + \alpha_3 X_{1i}^2 + \alpha_4 X_{2i}^2}},$$

where $(\beta_0, \beta_1, \beta_2, \beta_3, \beta_4, \beta_5) = (.5, 2, 1, 1, .4, .6)$ and $(\alpha_0, \alpha_1, \alpha_2, \alpha_3, \alpha_4) = (.1, .7, .5, -.4, .2)$. In this case, the average treatment effect is now equal to $\tau = \beta_1 + \beta_4 E(X_1) + \beta_5 E(X_2)$ $= 2.12$.

Table 3.8 and Figure 3.9 present results for the case where $N = 50$. The optimal Known Propensity Score GMM estimator is the $P = 1$ estimator. The optimal Unknown Propensity Score GMM estimator is also the $P = 1$ estimator. The optimal HIR estimator uses cross-validation for bandwidth selection. Notice in the tables to follow that Silverman's Rule of Thumb has a 39% – 92% higher MSE than cross-validation throughout all of the vector X results. Except for one case in the previous section where cross-validation outperformed Silverman's Rule of Thumb in a scalar setting, the results conform perfectly to the fact that Silverman's Rule of Thumb has nice optimality properties in one dimension, but not in higher dimensions. Figure 3.9 shows that $\bar{S}(P)$ again does a poor job in small samples of selecting the optimal P for the Unknown Propensity Score GMM estimator. The mode of the PDF of $\bar{P}$ is at $P = 0$ and $P = 1$ is the optimal P value. Table 3.8 shows that the optimal Known Propensity Score GMM estimator again has the lowest MSE, followed again by the optimal Unknown Propensity Score GMM estimator, and then by the optimal HIR estimator. The optimal Known Propensity Score GMM estimator has a 14.0% MSE reduction relative to the optimal Unknown Propensity Score GMM estimator and a 17.7% MSE reduction relative to the optimal HIR estimator. Also, the Unknown Propensity Score GMM estimator has a 4.3% MSE reduction relative to the HIR estimator.

Table 3.9 and Figure 3.10 present results for the case where sample size is increased to $N = 150$. In this case, the PDF of $\tilde{P}$ is still thick, but it is now centered around a mode at the true optimal P value, $P = 1$. The MSE ordering of the three estimators remains the same and the percent MSE difference between the optimal Known Propensity Score GMM estimator and the optimal Unknown Propensity Score GMM estimator decreases. The percent MSE difference between the optimal HIR estimator and the two GMM estimators is actually higher. This is most likely because the same cross-validation candidate bandwidth vector that was used when $N = 50$ was also used in this case with $N = 150$. The candidate bandwidth vector is more finely refined for the $N = 500$ simulations. This candidate bandwidth rule was also used with the one-dimensional artificial data, but it did not adversely affect the results. This shows that in higher dimensions the MSE of the cross-validation estimators can be very sensitive to bandwidth selection.

Table 3.10 and Figure 3.11 show results for the case where sample size is increased $N = 500$. The optimal Known Propensity Score GMM estimator has increased to the $P = 4$ estimator and the optimal Unknown Propensity Score GMM estimator has increased to the $P = 2$ estimator. The optimal HIR estimator still uses cross-validation. The percentage MSE reductions between the three estimators all now lie between 1.1% and 5.4%. Also, the Unknown Propensity Score GMM estimator is converging to the Known Propensity Score GMM estimator faster than the HIR estimator. Figure 3.11 shows that the PDF of $\tilde{P}$ is centered around the true optimal P value, $P = 2$, however the distribution is not as thin in the tails as with the same sample size in the one-dimensional case. Notice that the MSE of the $P = 0$ estimator is over 17 times the MSE of the optimal $P = 2$ estimator and yet the selection criterion function chooses $\tilde{P} = 0$ 17% of the time. This is further evidence that the $\tilde{S}(P)$ selection criterion is flawed (which is known) but the Unknown Propensity Score GMM estimator is more efficient than the HIR estimator in finite samples *assuming* the optimal P or K value can be chosen.

3.4.2 Monte Carlo Simulations Calibrated to LaLonde Data

A vector X Monte Carlo dataset is calibrated to the LaLonde dataset in a similar fashion to Section 3.3.2. The observed covariates are now the vector $(x_{L1}, x_{L2})'$. To calibrate the true propensity score, a Logit regression of t_L on $\left(1, x_{L1}, x_{L2}, x_{L1}^{\ 2}, x_{L2}^{\ 2}\right)$ is run and the resulting estimates obtained $\hat{\alpha} = (\hat{\alpha}_0, \hat{\alpha}_1, \hat{\alpha}_2, \hat{\alpha}_4, \hat{\alpha}_5)'$. The last difference is that an OLS regression of

$$y_L = \beta_0 + \beta_1 t_L + \beta_2 x_{L1} + \beta_3 x_{L2} + \beta_4 t_L x_{L1} + \beta_5 t_L x_{L2} + \varepsilon,$$

is run to estimate $\hat{\beta}$ and $\hat{\sigma}^2 = \hat{\varepsilon}'\hat{\varepsilon} / (N-6)$ for the calibrated outcome equation. All other components of the calibrated simulation routine remain the same.

It was discussed in the previous section that the calibrated true propensity score values were close enough to zero, by empirical standards, such that there were many singularities and the estimates were unreliable and volatile. Therefore, a truncation was made to the LaLonde data and one of the propensity score parameters was adjusted in order to center the distribution of the true propensity score. For the same reasons, similar actions are employed in the vector X case. Since, of the three non-indicator variables, Age and Education have the best overlap in their distributions between the treated and the controls, they were selected as the two covariates for use in this section. Given the use of these two variables in the calibration routine described above, the minimum pre-modification true propensity score value was .0000256, the maximum value was .590, and the mean value was .069. The minimum value is definitely not empirically suitable for average treatment effect estimation. Thus, the following changes were made. Observations with an Age value less than 17, an Age value greater than 48, an Education value less than 8, or an Education value greater than 12 were deleted from the LaLonde dataset. This resulted in a decrease in sample size from 2675 to 1406. Again, with only this adjustment, the propensity score remained low so the $\hat{\alpha}_1$ parameter was decreased to 45% of its original value. The new true propensity score has minimum value of .173, a maximum value of .853, and a mean value of .457. The resulting distribution of the true propensity score is more centered and bounded away from zero and one. Figure 3.12 shows the histogram of the original pre-modification propensity score and the histogram of the post-modification propensity score. Finally, to better enable higher-order polynomials to be employed in the unknown propensity score routine, sample sizes of 100, 250, and 500 are used.

Table 3.11 and Figure 3.13 present results for the LaLonde case where $N = 100$. Sample size is small enough such that the optimal polynomial order is $P = 0$ for both GMM estimators. The optimal HIR estimator still uses cross-validation for bandwidth selection and will use cross-validation for the remaining sample sizes as well. Figure 3.13 shows that the true optimal P value for the Unknown Propensity Score GMM estimator is selected almost 80% of the time by the selection criterion. Also, since the true optimal polynomial order is $P = 0$, the HIR cross-validation estimator has a lower MSE than the Unknown Propensity Score GMM estimator. However, knowledge of the propensity score reduces this MSE by 3.9%.

Table 3.12 and Figure 3.14 show results when sample size is $N = 250$. Here sample size has increased enough such that the optimal Unknown Propensity Score GMM estimator is no longer $P = 0$, but $P = 1$. Hence, as one would predict, the optimal Unknown Propensity Score GMM estimator performs better than the optimal HIR estimator. The optimal Known Propensity Score GMM estimator still outperforms the other two. Another interesting result is that the PDF $\bar{P}$ shows that $\tilde{S}(P)$ performs poorly again and overwhelmingly selects the sub-optimal $P = 0$ estimator.

Table 3.13 and Figure 3.15 show results for the final case where sample size is increased to $N = 500$. The standard MSE ordering of the three estimators can be seen. Also, in MSE terms, all three estimators appear to be converging together as sample size gets larger. Finally, the $\tilde{S}(P)$ selection criterion still performs poorly. One would expect the PDF of $\bar{P}$ to converge more slowly to a tight distribution around the optimal P as the dimensionality of X increases, as in the artificial simulation results of Section 3.4.1. However, in this case, the PDF appears to be converging not slowly to the optimal P value, but slowly to a sub-optimal one.

3.5 Conclusions

This chapter started by defining the adjusted GMM framework for use in estimation of average treatment effects when the propensity score is unknown. The framework essentially estimates the propensity score with a Logit regression of the treatment indicator on higher-order polynomials of X. The fitted probability function is plugged into the weighting function, $\tilde{V}$, to obtain the average treatment effect estimate. The key difference between this method of estimating an average treatment effect and the HIR method is that HIR use a non-parametric estimate of the propensity score and this method uses a flexible Logit specification. Given this specification, a criterion was developed for selection of K or P. This selection criterion was simplified to a very practical and easily calculable quantity that ignored lower-order covariance terms in the true MSE of the resulting estimator, $\tilde{\tau}$.

Monte Carlo simulations were performed on artificial data and on data calibrated to LaLonde's 1986 non-experimental dataset so that the finite sample properties of the Known Propensity Score GMM estimator, the Unknown Propensity Score GMM estimator, and the HIR estimator could be examined. The LaLonde dataset upon which the calibration was based had to be adjusted so that the true propensity score (known to the simulator, not to the econometrician) was significantly centered in the interval $(0,1)$ for all practical purposes.

The simulations confirm the finite sample intuition being developed in this dissertation. The Known Propensity Score GMM estimator has the lowest MSE relative to the two unknown propensity score estimators in every data generating process considered. The percent MSE reductions are as high as 14% – 17% when two covariates are used and sample size is small. As sample size increases, these MSE reductions are converging to zero. This analysis serves as the finite sample compliment to the asymptotic results of Hahn (1998). Asymptotically, the propensity score is ancillary for estimation of average treatment effects. However, in finite samples, the propensity score is *not ancillary* for estimation of average treatment effects, based upon the results relative to two alternative asymptotically efficient unknown propensity score estimators.

The optimal Unknown Propensity Score GMM estimator has a lower MSE than the optimal HIR estimator in all cases but two. The two cases are when the optimal order of X is $K=0$ or $P=0$. In this case X is weak and it is necessarily true that the non-parametric cross-validation estimator will outperform what reduces down to the simple difference-in-averages estimator. In all other cases considered, the Unknown Propensity Score GMM estimator outperforms the HIR estimator by as much as 7%. Thus, in finite samples even if the propensity score is unknown, there can be a non-trivial efficiency gain, relative to the HIR estimator, from optimal inclusion of a subset of the information contained in X. Similar results also obtain under various other data generating processes and with use of the Linear Probability Model instead of the Logit. However, estimation of the propensity score with the Linear Probability Model performed slightly worse than with the Logit.

For the Unknown Propensity Score GMM estimator to be feasible in practice, there must exist an effective selection criterion upon which moment (polynomial) selection can be based. The higher-order asymptotically optimal moment selection criterion of Chapters 1 and 2, $\hat{S}(K)$, has been shown to perform well in practice, even in small samples. The selection criterion developed in this chapter, $\bar{S}(K)$, is not asymptotically optimal. It was chosen for its simplicity. The simulations confirm that it tends to perform poorly in small samples and performs better as sample size becomes large and the effect of the omitted lower-order covariance terms converges to zero. However, as sample size becomes large the efficiency gain of the Unknown Propensity Score GMM estimator relative to the HIR estimator diminishes. If sample size is large enough for one to feel comfortable in using $\bar{S}(K)$, one might as well use the HIR estimator with cross-validation or Silverman's Rule of Thumb. Silverman's Rule of Thumb tends to be optimal with a scalar covariate and cross-validation tends to be optimal with a vector of covariates. Thus, until an alternate selection criterion is developed, the advantages of the Unknown Propensity Score GMM estimator over the HIR estimator in finite sample estimation of average treatment effects may go unused.

This dissertation has served to develop and extend the field's understanding of estimation of average treatment effects with finite samples. The motivation behind this is simply that many experimental datasets are often very small and some non-experimental datasets are what would be classified as small. Thus, the asymptotic properties of many commonly used estimators will not hold in practice. This dissertation has shown the finite sample advantages of two GMM estimators of average treatment effects, one when the propensity score is known and the other when it is unknown. The selection criterion for the Known Propensity Score GMM estimator is higher-order asymptotically optimal and performs very well in practice. The selection criterion for the Unknown Propensity Score GMM estimator is not asymptotically optimal and is not very reliable in practice. Some possible extensions of the analysis of this dissertation would be to derive an asymptotically optimal selection criterion for the Unknown Propensity Score GMM estimator and examine its finite sample performance. Also, calibration with the unmodified LaLonde non-experimental data was not possible due to the closeness of the "true" propensity score to zero. Empirical examination of the range in which Assumption 4 is valid and not valid would be of value to researchers who typically use datasets with a significantly disproportionate number of units of treated or controls.

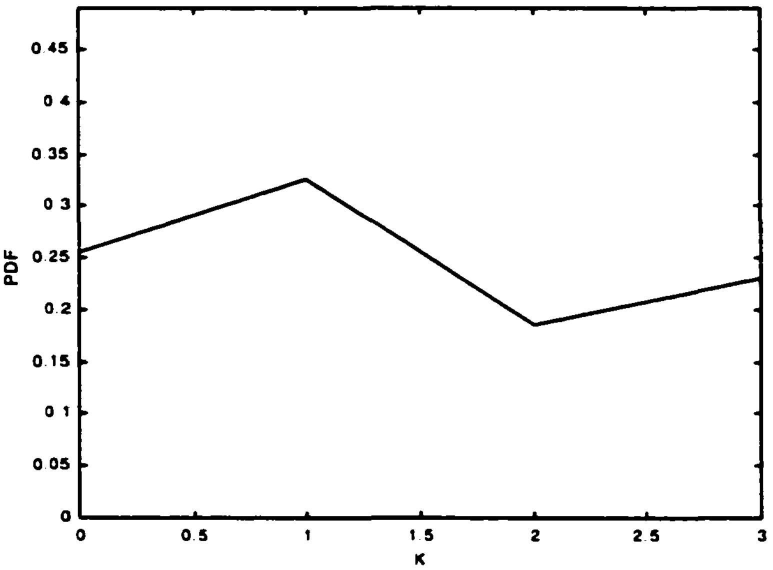

Figure 3.1: PDF of $\tilde{K}$ Over 1000 Samples.

Monte Carlo Simulations with Fake Data and One Covariate. (N = 50)

	$p(x)$ Known	$p(x)$ Unknown	
K	GMM Estimator	GMM Estimator	Non-Parametric Estimator
-1	.2782		
0	.1059	.2306	Cross-Validation
1	***.0923****	***.0960****	.1000
2	.0929	.0985	
3	.0951	.1031	Silverman's Rule of Thumb
4	.1001		***.0982****
5	.1012		
6	.1021		

Table 3.1: MSE of Various Estimators When the Propensity Score is Known and Unknown.

Monte Carlo Simulations with Fake Data and One Covariate. (N = 50)

Optimal GMM Estimator of Each Class in Bold and Italics with an Asterisk.

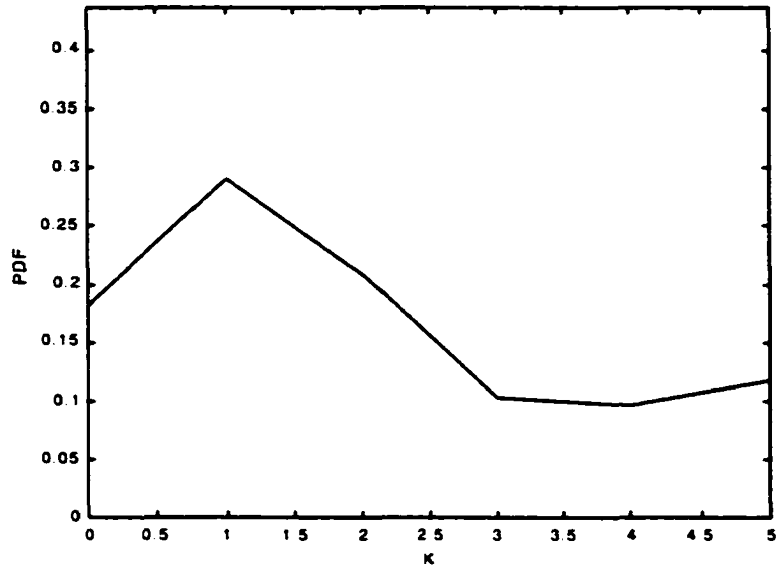

Figure 3.2: PDF of $\bar{K}$ Over 1000 Samples.

Monte Carlo Simulations with Fake Data and One Covariate. (N = 150)

	$p(x)$ Known	$p(x)$ Unknown	
K	GMM Estimator	GMM Estimator	Non-Parametric Estimator
-1	.0974		Cross-Validation .0316 Silverman's Rule of Thumb ***.0306****
0	.0351	.1516	
1	.0297	***.0302****	
2	.0295	.0304	
3	***.0292****	.0308	
4	.0303	.0305	
5	.0301	.0315	
6	.0303		

Table 3.2: MSE of Various Estimators When the Propensity Score is Known and Unknown.

Monte Carlo Simulations with Fake Data and One Covariate. (N = 150)

Optimal GMM Estimator of Each Class in Bold and Italics with an Asterisk.

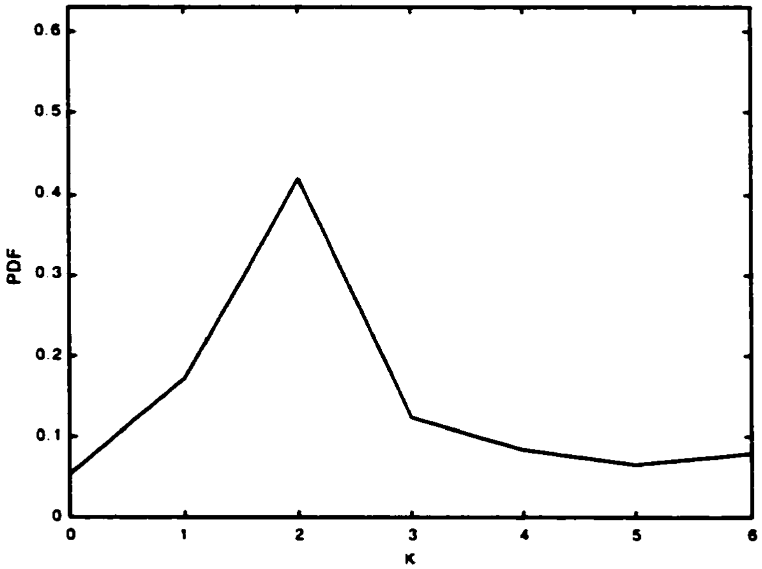

Figure 3.3: PDF of $\tilde{K}$ Over 1000 Samples.

Monte Carlo Simulations with Fake Data and One Covariate. (N = 500)

	$p(x)$ Known	$p(x)$ Unknown	
K	GMM Estimator	GMM Estimator	Non-Parametric Estimator
-1	.02894		
0	.01030	.12627	Cross-Validation
1	.00882	.00878	.00993
2	.00862	***.00867****	
3	***.00862****	.00890	Silverman's Rule of Thumb
4	.00885	.00890	***.00883****
5	.00902	.00887	
6	.00896	.00907	

Table 3.3: MSE of Various Estimators When the Propensity Score is Known and Unknown.

Monte Carlo Simulations with Fake Data and One Covariate. (N = 500)

Optimal GMM Estimator of Each Class in Bold and Italics with an Asterisk.

	Full Sample	Treated (NSW)	Controls (PSID)
Earnings 1978 (Y)	20502 (15633)	6349 (7867)	21554 (15555)
Treatment Status (T)	.069 (.254)	1 (0)	0 (0)
Earnings 1975	17851 (13878)	1532 (3219)	19063 (13597)
Age	34.2 (10.5)	25.8 (7.2)	34.9 (10.4)
Education	12.0 (3.1)	10.4 (2.0)	12.1 (3.1)
Married	.819 (.385)	.189 (.393)	.866 (.340)
Black	.292 (.455)	.843 (.365)	.251 (.334)
Hispanic	.034 (.182)	.060 (.237)	.033 (.177)

Table 3.4: Sample Means (Standard Deviation) of Post-Training Earnings, Treatment Status, and Pre-Treatment Variables for the Treated Sub-Sample of the NSW AFDC Participants and the PSID Controls of LaLonde (1986).

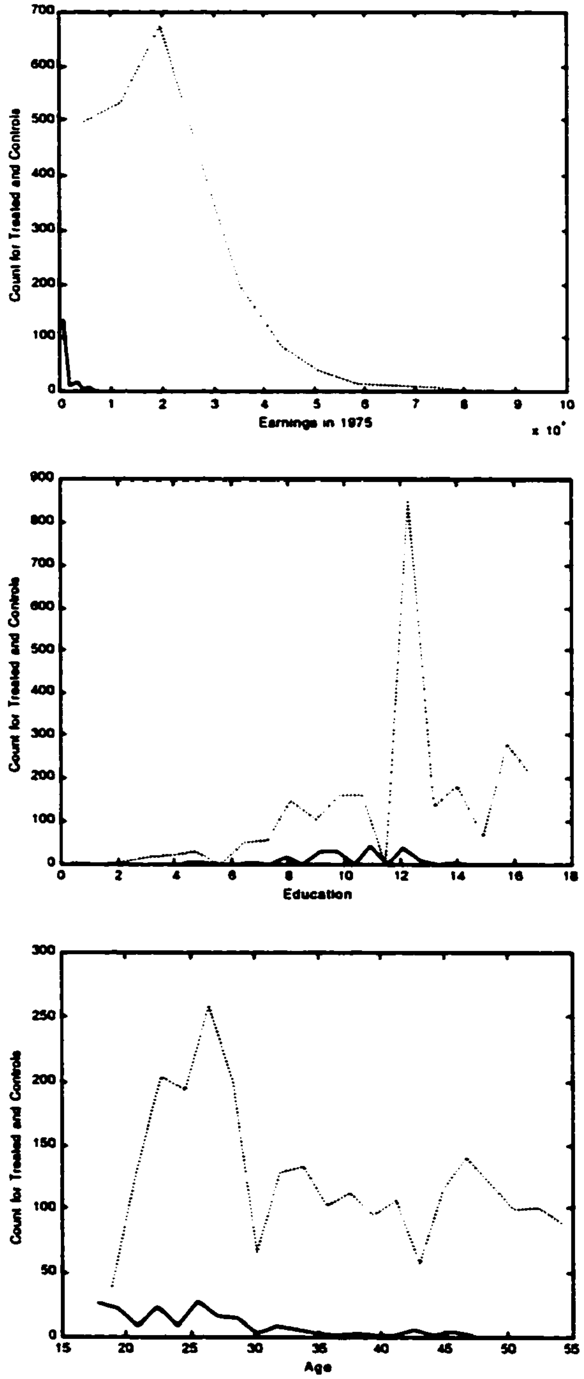

Figure 3.4: Histogram Plot of Earnings in 1975 (Top), Education (Middle), and Age (Bottom) for Treated (Solid) and Controls (Dotted).

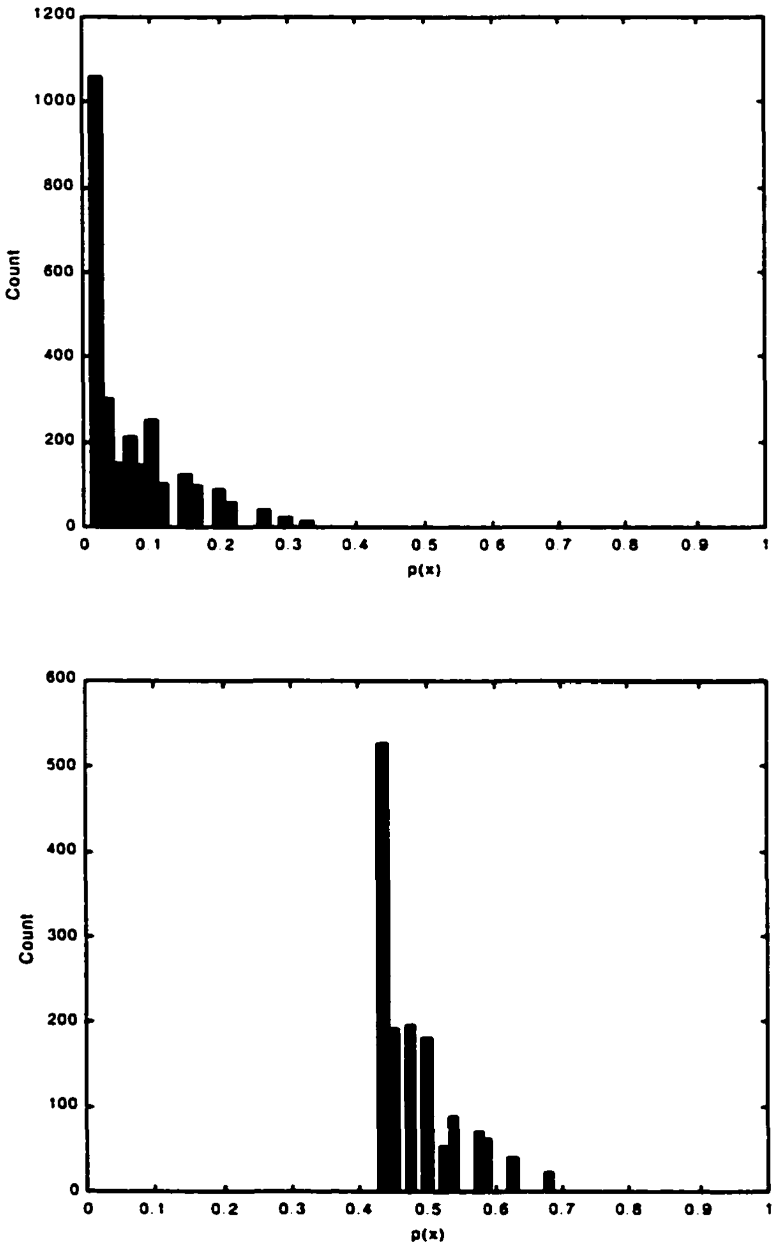

Figure 3.5: Histogram of True Propensity Score Values Before Modification (Top) and After Modification (Bottom).

(X = Age)

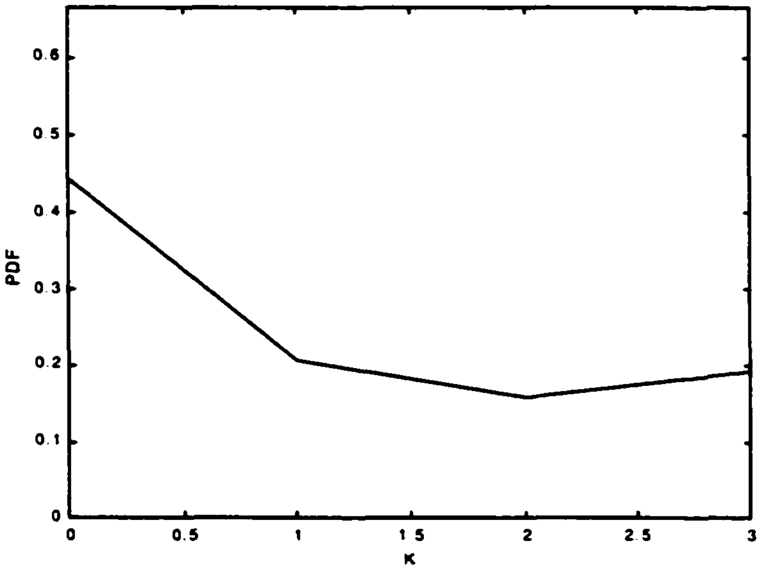

Figure 3.6: PDF of $\tilde{K}$ Selection Over 1000 Samples.

DGP Calibrated to LaLonde's Non-Experimental Data with One Covariate. (N = 50)

	$p(x)$ Known	$p(x)$ Unknown	
K	GMM Estimator	GMM Estimator	Non-Parametric Estimator
-1	2.925		Cross-Validation ***1.425**** Silverman's Rule of Thumb 1.527
0	1.422	***1.458****	
1	***1.386****	1.466	
2	1.457	1.498	
3	1.498	1.538	
4	1.531		
5			

Table 3.5: MSE of Various Estimators When the Propensity Score is Known and Unknown.

DGP Calibrated to LaLonde's Non-Experimental Data with One Covariate. (N = 50)

Optimal GMM Estimator of Each Class in Bold and Italics with an Asterisk.

$(\times 10^7)$

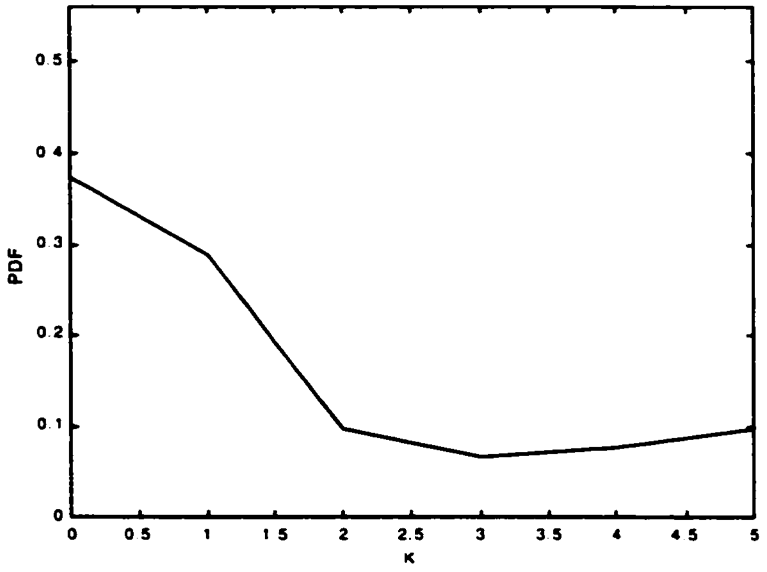

Figure 3.7: PDF of $\tilde{K}$ Over 1000 Samples.

DGP Calibrated to LaLonde's Non-Experimental Data with One Covariate. (N =150)

	$p(x)$ Known	$p(x)$ Unknown	
K	GMM Estimator	GMM Estimator	Non-Parametric Estimator
-1	9.681		
0	4.763	4.782	Cross-Validation
1	**_4.514_***	**_4.656_***	4.723
2	4.638	4.707	
3	4.666	4.733	Silverman's Rule of Thumb
4	4.777	4.833	**_4.711_***
5	4.786	4.968	

Table 3.6: MSE of Various Estimators When the Propensity Score is Known and Unknown.

DGP Calibrated to LaLonde's Non-Experimental Data with One Covariate. (N = 150)

Optimal GMM Estimator of Each Class in Bold and Italics with an Asterisk.

$(\times 10^6)$

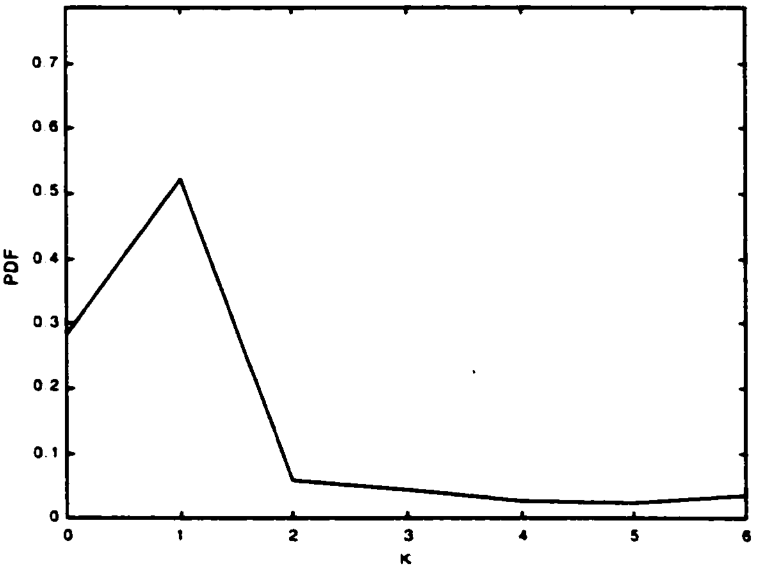

Figure 3.8: PDF of $\bar{K}$ Over 1000 Samples.

DGP Calibrated to LaLonde's Non-Experimental Data with One Covariate. (N = 500)

	$p(x)$ Known	$p(x)$ Unknown	
K	GMM Estimator	GMM Estimator	Non-Parametric Estimator
-1	2.924		
0	1.427	1.423	<u>Cross-Validation</u>
1	1.394	***1.368****	1.484
2	***1.325****	1.385	
3	1.343	1.386	<u>Silverman's Rule of Thumb</u>
4	1.376	1.406	***1.393****
5	1.402	1.420	
6	1.434	1.444	

Table 3.7: MSE of Various Estimators When the Propensity Score is Known and Unknown.

DGP Calibrated to LaLonde's Non-Experimental Data with One Covariate. (N = 500)

Optimal GMM Estimator of Each Class in Bold and Italics with an Asterisk.

$\left(\times 10^{6}\right)$

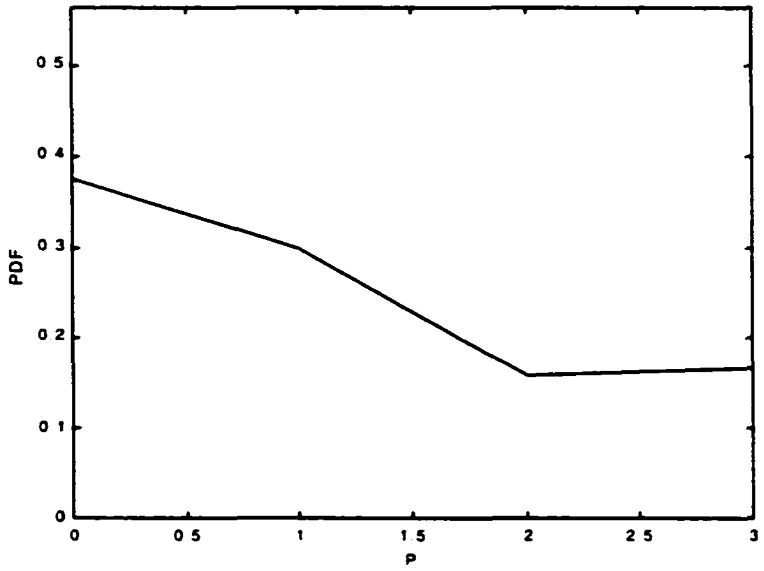

Figure 3.9: PDF of $\bar{P}$ Over 1000 Samples.

Monte Carlo Simulations with Fake Data and Two Covariates. (N = 50)

	$p(x)$ Known	$p(x)$ Unknown	
P	GMM Estimator	GMM Estimator	Non-Parametric Estimator
-1	.373		
0	.129	.287	Cross-Validation
1	***.0922****	***.107****	***.112****
2	.100	.127	
3	.104	.146	Silverman's Rule of Thumb
4	.114		.215
5			

Table 3.8: MSE of Various Estimators When the Propensity Score is Known and Unknown.

Monte Carlo Simulations with Fake Data and Two Covariates. (N = 50)

Optimal GMM Estimator of Each Class in Bold and Italics with an Asterisk.

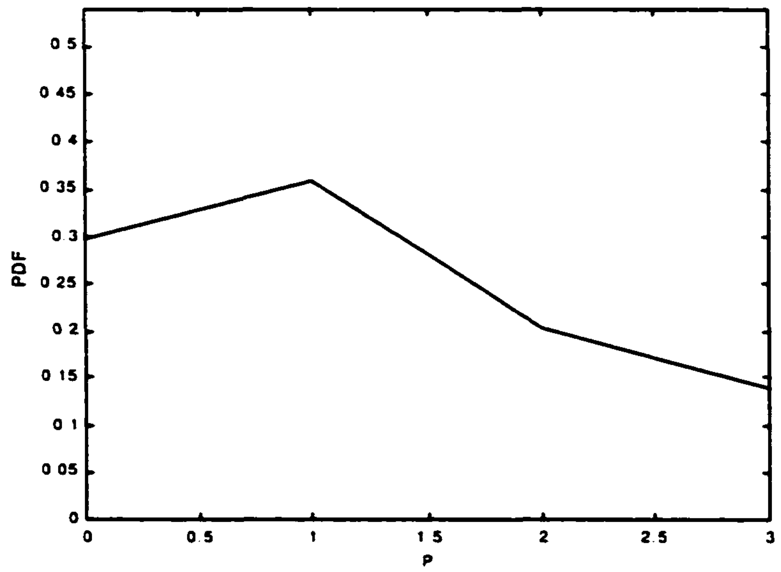

Figure 3.10: PDF of $\tilde{P}$ Over 1000 Samples.

Monte Carlo Simulations with Fake Data and Two Covariates. (N = 150)

	$p(x)$ Known	$p(x)$ Unknown	
P	GMM Estimator	GMM Estimator	Non-Parametric Estimator
-1	.1172		
0	.0435	.1943	<u>Cross-Validation</u>
1	.0298	***.0307****	***.0377****
2	***.0294****	.0316	
3	.0313	.0340	<u>Silverman's Rule of Thumb</u>
4	.0316		.0570
5	.0327		
6	.0332		

Table 3.9: MSE of Various Estimators When the Propensity Score is Known and Unknown.

Monte Carlo Simulations with Fake Data and Two Covariates. (N = 150)

Optimal GMM Estimator of Each Class in Bold and Italics with an Asterisk.

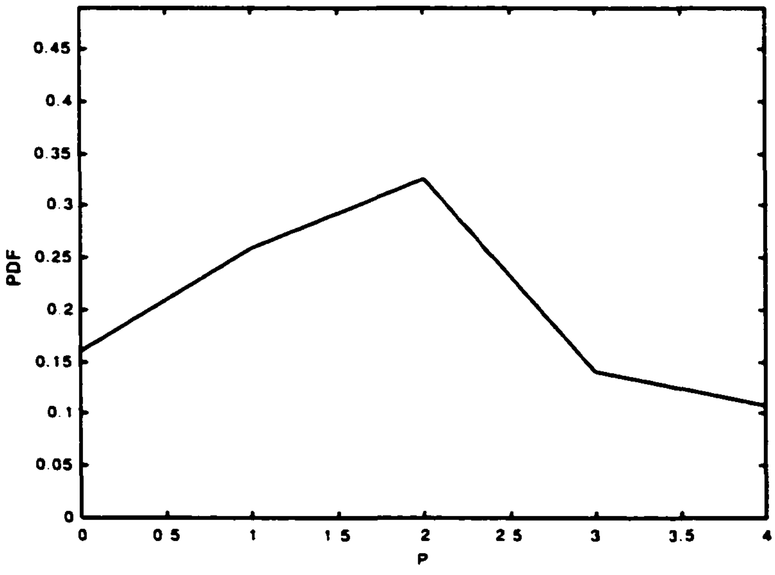

Figure 3.11: PDF of $\bar{P}$ Over 1000 Samples.

Monte Carlo Simulations with Fake Data and Two Covariates. (N = 500)

	$p(x)$ Known	$p(x)$ Unknown	
P	GMM Estimator	GMM Estimator	Non-Parametric Estimator
-1	.03599		
0	.01298	.16131	Cross-Validation
1	.00900	.00886	***.00916****
2	.00896	***.00876****	
3	.00886	.00898	Silverman's Rule of Thumb
4	***.00867****	.00898	.01400
5	.00905		
6	.00911		
7	.00917		
8	.00929		

Table 3.10: MSE of Various Estimators When the Propensity Score is Known and Unknown.

Monte Carlo Simulations with Fake Data and Two Covariates. (N = 500)

Optimal GMM Estimator of Each Class in Bold and Italics with an Asterisk.

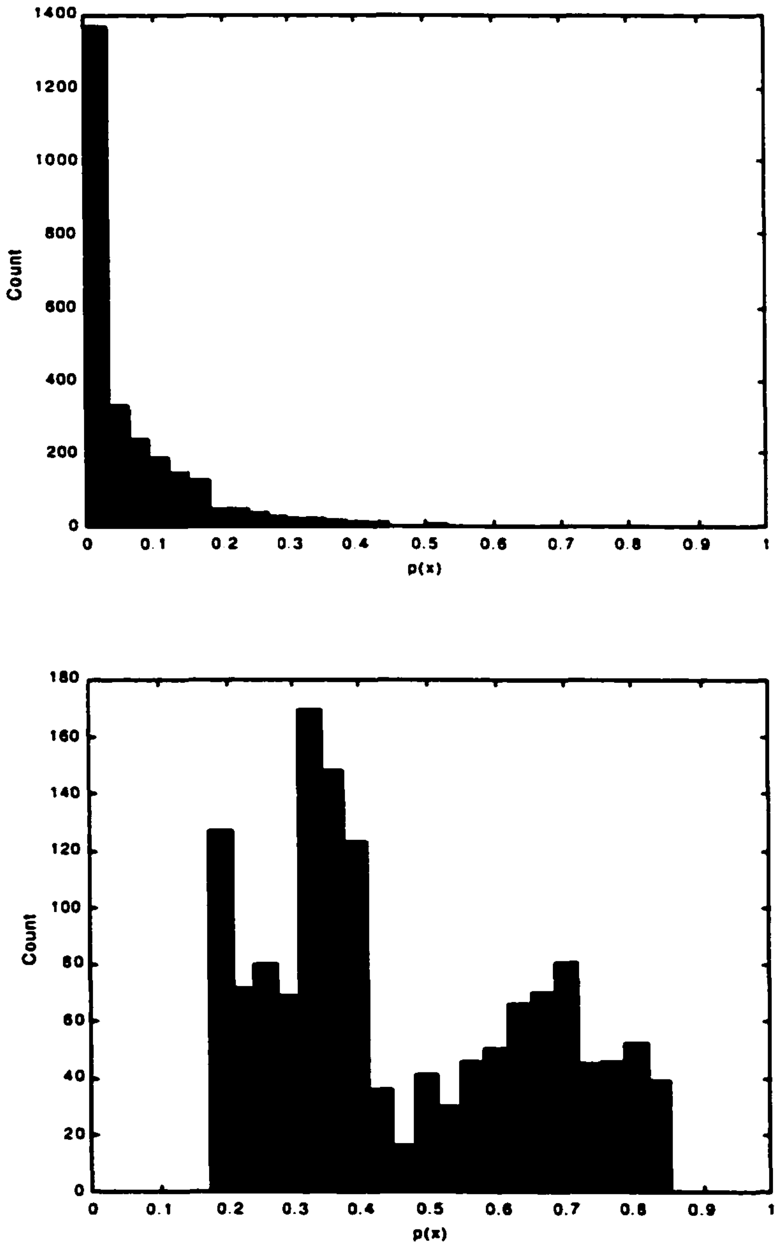

Figure 3.12: Histogram of True Propensity Score Values Before Modification (Top) and After Modification (Bottom).

(X = Age and Education)

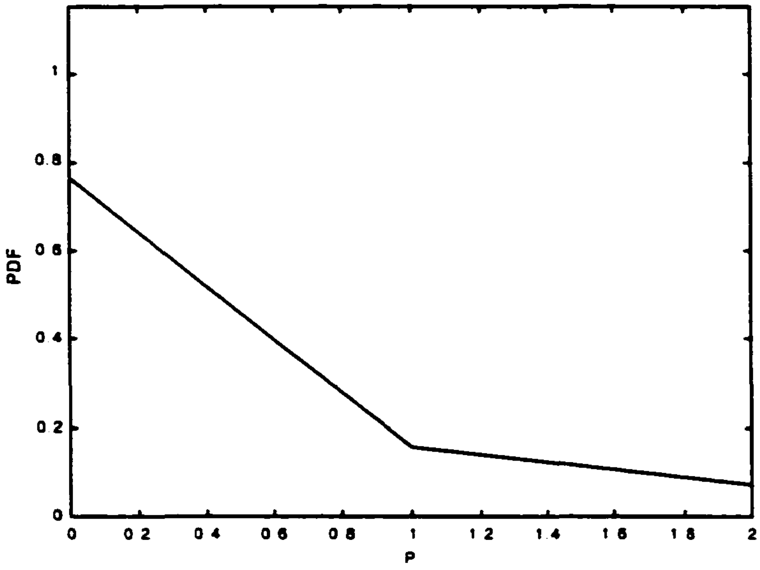

Figure 3.13: PDF of $\tilde{P}$ Over 1000 Samples.

DGP Calibrated to LaLonde's Non-Experimental Data with Two Covariates. (N = 100)

	$p(x)$ Known	$p(x)$ Unknown	
P	GMM Estimator	GMM Estimator	Non-Parametric Estimator
-1	13.994		
0	***7.049****	***7.674****	Cross-Validation
1	7.450	8.597	***7.331****
2	7.672	8.663	
3	8.043		Silverman's Rule of Thumb
4	8.371		9.630
5			

Table 3.11: MSE of Various Estimators When the Propensity Score is Known and Unknown.

DGP Calibrated to LaLonde's Non-Experimental Data with Two Covariates. (N = 100)

Optimal GMM Estimator of Each Class in Bold and Italics with an Asterisk.

$(\times 10^6)$

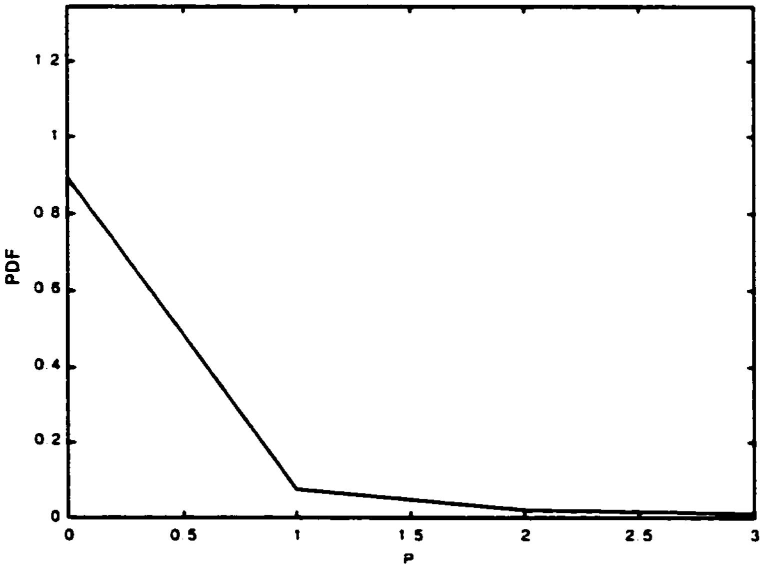

Figure 3.14: PDF of $\bar{P}$ Over 1000 Samples.

DGP Calibrated to LaLonde's Non-Experimental Data with Two Covariates. (N = 250)

	$p(x)$ Known	$p(x)$ Unknown	
P	GMM Estimator	GMM Estimator	Non-Parametric Estimator
-1	5.597		
0	2.883	3.854	Cross-Validation
1	***2.876****	***3.072****	***3.294****
2	3.003	3.105	
3	3.033	3.177	Silverman's Rule of Thumb
4	3.056		4.227
5			

Table 3.12: MSE of Various Estimators When the Propensity Score is Known and Unknown.

DGP Calibrated to LaLonde's Non-Experimental Data with Two Covariates. (N = 250)

Optimal GMM Estimator of Each Class in Bold and Italics with an Asterisk.

$(\times 10^6)$

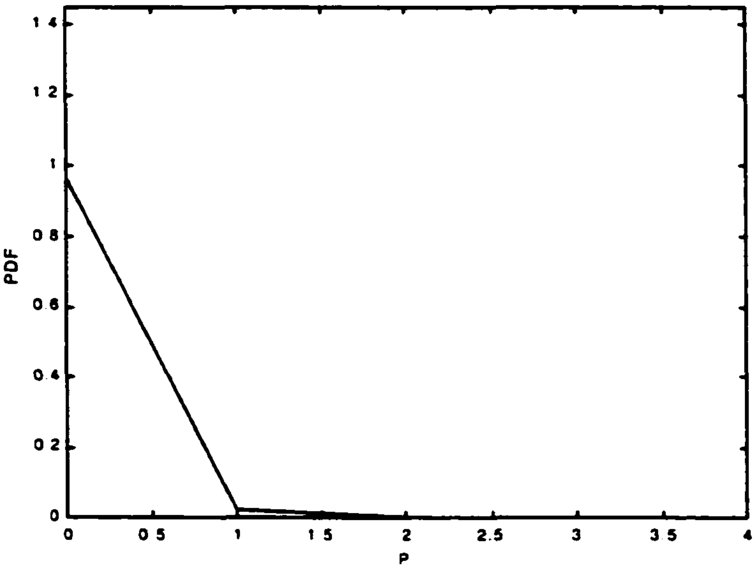

Figure 3.15: PDF of $\tilde{P}$ Over 1000 Samples.

DGP Calibrated to LaLonde's Non-Experimental Data with Two Covariates. (N = 500)

	$p(x)$ Known	$p(x)$ Unknown	
P	GMM Estimator	GMM Estimator	Non-Parametric Estimator
-1	2.716		
0	1.462	2.781	Cross-Validation
1	1.421	1.568	***1.493*** *
2	***1.420*** *	***1.477*** *	
3	1.492	1.509	Silverman's Rule of Thumb
4	1.481	1.547	2.074
5			

Table 3.13: MSE of Various Estimators When the Propensity Score is Known and Unknown.

DGP Calibrated to LaLonde's Non-Experimental Data with Two Covariates. (N = 500)

Optimal GMM Estimator of Each Class in Bold and Italics with an Asterisk.

$(\times 10^6)$

4. Appendix

Proof of Proposition 1.1:

Let X be discrete with known support $(\gamma_1, \gamma_2, \ldots \gamma_J)$, $p(X=\gamma_j)=$ $p(T=1 \mid X=\gamma_j)=\mu_j$, and $(\pi_1, \pi_2, \ldots, \pi_N)$ be a vector of nuisance parameters. Also, let N_{tj} be the number of observations with $t_i = t$ and $x_i = \gamma_j$, and let $N_{\cdot j}$ be the number of observations with $x_i = \gamma_j$. The discrete analog of the GMM framework is

$$\psi_1(y,t,x,\tau)=\frac{y \cdot t}{p(x)}-\frac{y \cdot (1-t)}{1-p(x)}-\tau,$$

$$\psi_2(t,x)=\begin{pmatrix} 1\{x=\gamma_1\}(t-p(x)) \\ 1\{x=\gamma_2\}(t-p(x)) \\ \vdots \\ 1\{x=\gamma_J\}(t-p(x)) \end{pmatrix}.$$

Thus, ψ contains $M=J+1$ moments. The ψ_1 moment has remained unchanged and a linear combination of the ψ_2 moments perfectly accounts for $a(x)(t-p(x))$ for any a. Given this setup, the optimization problem of the Empirical Likelihood estimator is

$$\max_{\pi_1,\ldots \pi_N, \tau} L(\pi)=\sum_{i=1}^{N} \ln \pi_i \quad \text{s.t. (i)} \sum_{i=1}^{N} \pi_i = 1$$

$$\text{(ii)} \sum_{i=1}^{N} \pi_i \left(\frac{y_i \cdot t_i}{p(x_i)} - \frac{y_i \cdot (1-t_i)}{1-p(x_i)} - \tau \right) = 0$$

$$\text{(iii)} \sum_{i=1}^{N} \pi_i \left(1\{x_i = \gamma_j\} \left(t_i - p(x_i) \right) \right) = 0 \quad \forall j \in \{1, \ldots, J\},$$

where $1\{B\}$ is an indicator function equal to one if B is true and zero otherwise.

Therefore, the Lagrangian is

$$\mathcal{L} = \sum_{i=1}^{N} \ln \pi_i + \eta \left(1 - \sum_{i=1}^{N} \pi_i \right) + \sum_{j=1}^{J} \lambda_j \left(- \sum_{i=1}^{N} \pi_i \left(1\{x_i = \gamma_j\} \left(t_i - p(x_i) \right) \right) \right).$$

Note that

$$\hat{\tau} = \sum_{i=1}^{N} \hat{\pi}_i \left(\frac{y_i \cdot t_i}{p(x_i)} - \frac{y_i \cdot (1 - t_i)}{1 - p(x_i)} \right)$$

and its Lagrange Multiplier is identically zero because relaxing the (ii) restriction changes the value of $\hat{\tau}$ but not the maximized value of $L(\pi)$. Taking the derivative with respect to π_i yields

$$\frac{1}{\pi_i} - \eta - \sum_{j=1}^{J} \lambda_j \left(1\{x_i = \gamma_j\} \left(t_i - p(x_i) \right) \right) = 0 \quad \forall i \in \{1, \ldots, N\}.$$

Multiplying each side by π_i and summing over all N equations yields $\eta = N$. Solving for π_i gives

$$\hat{\pi}_i = \frac{\frac{1}{N}}{1 + \sum_{j=1}^{J} \frac{\lambda_j}{N} \left(1\{x_i = \gamma_j\} \left(t_i - p(x_i) \right) \right)} \quad \forall i \in \{1, \ldots, N\}.$$

Combining $\hat{\pi}_i$ with (iii) implies

$$\frac{1}{N} \sum_{i=1}^{N} \frac{1\{x_i = \gamma_j\} \left(t_i - p(x_i) \right)}{1 + \sum_{r=1}^{J} \frac{\lambda_r}{N} \left(1\{x_i = \gamma_r\} \left(t_i - p(x_i) \right) \right)} = 0 \quad \forall j \in \{1, \ldots, J\}.$$

For a given j, this implies

$$\sum_{i|x_i=\gamma_j} \frac{(t_i-p(x_i))}{1+\frac{\lambda_j}{N}(t_i-p(x_i))}=0$$

$$\Rightarrow N_{1j}\frac{(1-\mu_j)}{1+\frac{\lambda_j}{N}(1-\mu_j)}+(N_{\cdot j}-N_{1j})\frac{-\mu_j}{1-\frac{\lambda_j}{N}\mu_j}=0$$

$$\Rightarrow N_{1j}\frac{(1-\mu_j)}{1+\frac{\lambda_j}{N}(1-\mu_j)}=(N_{\cdot j}-N_{1j})\frac{\mu_j}{1-\frac{\lambda_j}{N}\mu_j}$$

$$\Rightarrow N_{1j}(1-\mu_j)-N_{1j}(1-\mu_j)\frac{\lambda_j}{N}\mu_j=(N_{\cdot j}-N_{1j})\mu_j+(N_{\cdot j}-N_{1j})\mu_j\frac{\lambda_j}{N}(1-\mu_j)$$

$$\Rightarrow N_{1j}=N_{\cdot j}\mu_j+N_{\cdot j}\mu_j(1-\mu_j)\frac{\lambda_j}{N}$$

$$\Rightarrow \frac{\lambda_j}{N}=\frac{N_{1j}-N_{\cdot j}\mu_j}{N_{\cdot j}\mu_j(1-\mu_j)}=\frac{\frac{N_{1j}}{N_{\cdot j}}-\mu_j}{\mu_j(1-\mu_j)} \quad \forall j\in\{1,\dots,J\}.$$

Thus, the Empirical Likelihood estimator of τ is

$$\hat{\tau}_{EL}=\sum_{i=1}^{N}\left(\frac{\frac{1}{N}}{1+\sum_{j=1}^{J}\frac{\lambda_j}{N}\left(1\{x_i=\gamma_j\}(t_i-p(x_i))\right)}\right)\left(\frac{y_i\cdot t_i}{p(x_i)}-\frac{y_i\cdot(1-t_i)}{1-p(x_i)}\right)$$

$$=\frac{1}{N}\sum_{i=1}^{N}\frac{\frac{y_i\cdot t_i}{p(x_i)}-\frac{y_i\cdot(1-t_i)}{1-p(x_i)}}{1+\sum_{j=1}^{J}\frac{\frac{N_{1j}}{N_{\cdot j}}-\mu_j}{\mu_j(1-\mu_j)}\left(1\{x_i=\gamma_j\}(t_i-p(x_i))\right)}.$$

Inspection will reveal that the Empirical Likelihood estimator above is equal to the non-parametric propensity score weighted estimator (denoted the "estimated weights" estimator in HIR)

$$\hat{\tau}_{HIR} = \frac{1}{N}\sum_{i=1}^{N}\frac{y_i \cdot t_i}{\hat{p}(x_i)} - \frac{y_i \cdot (1-t_i)}{1-\hat{p}(x_i)},$$

where $\hat{p}(x_i) = \sum_{j=1}^{J} 1\{x_i = \gamma_j\}\frac{N_{1j}}{N_{\cdot j}}$. Since the CUE is asymptotically equivalent to the EL estimator, it not only has the same asymptotic distribution as $\hat{\tau}_{HIR}$, but is also asymptotically identical to $\hat{\tau}_{HIR}$ under a general specification of the propensity score and discrete support X.

Q.E.D.

Proof of Asymptotic Expansion 1.1:

(1) Follows from the Central Limit Theorem.

(2) Follows from the Central Limit Theorem.

(3) $\left(\frac{1}{N}\sum_i \psi_{2i}\psi_{2i}'\right)^{-1}$

$$= \left(\Omega^{-1} - \Omega^{-1}\mathrm{H}\Omega^{-1}\right)\left(\Omega^{-1} - \Omega^{-1}\mathrm{H}\Omega^{-1}\right)^{-1}\left(\Omega + \mathrm{H}\right)^{-1}$$

$$= \left(\Omega^{-1} - \Omega^{-1}\mathrm{H}\Omega^{-1}\right)\left(\left(\Omega + \mathrm{H}\right)\left(\Omega^{-1} - \Omega^{-1}\mathrm{H}\Omega^{-1}\right)\right)^{-1}$$

$$= \left(\Omega^{-1} - \Omega^{-1}\mathrm{H}\Omega^{-1}\right)\left(I - \mathrm{H}\Omega^{-1} + \mathrm{H}\Omega^{-1} - \mathrm{H}\Omega^{-1}\mathrm{H}\Omega^{-1}\right)^{-1}$$

$$= \left(\Omega^{-1} - \Omega^{-1}\mathrm{H}\Omega^{-1}\right)\left(I - \mathrm{H}\Omega^{-1}\mathrm{H}\Omega^{-1}\right)^{-1}$$

$$=\left(\Omega^{-1}-\Omega^{-1}\mathrm{H}\Omega^{-1}\right)\left(I+\mathrm{H}\Omega^{-1}\mathrm{H}\Omega^{-1}\right)\left(I+\mathrm{H}\Omega^{-1}\mathrm{H}\Omega^{-1}\right)^{-1}\left(I-\mathrm{H}\Omega^{-1}\mathrm{H}\Omega^{-1}\right)^{-1}$$

$$=\left(\Omega^{-1}-\Omega^{-1}\mathrm{H}\Omega^{-1}\right)\left(I+\mathrm{H}\Omega^{-1}\mathrm{H}\Omega^{-1}\right)\left(\left(I-\mathrm{H}\Omega^{-1}\mathrm{H}\Omega^{-1}\right)\left(I+\mathrm{H}\Omega^{-1}\mathrm{H}\Omega^{-1}\right)\right)^{-1}$$

$$=\left(\Omega^{-1}-\Omega^{-1}\mathrm{H}\Omega^{-1}\right)\left(I+\mathrm{H}\Omega^{-1}\mathrm{H}\Omega^{-1}\right)\left(I-\mathrm{H}\Omega^{-1}\mathrm{H}\Omega^{-1}\mathrm{H}\Omega^{-1}\mathrm{H}\Omega^{-1}\right)^{-1}$$

$$=\left(\Omega^{-1}-\Omega^{-1}\mathrm{H}\Omega^{-1}\right)\left(I+\mathrm{H}\Omega^{-1}\mathrm{H}\Omega^{-1}\right)+o_p\left(N^{-1/2}\right)$$

$$=\Omega^{-1}-\Omega^{-1}\mathrm{H}\Omega^{-1}+\Omega^{-1}\mathrm{H}\Omega^{-1}\mathrm{H}\Omega^{-1}+o_p\left(N^{-1}\right)$$

In the proof of (3), we are keeping only the terms that are of order N^{-1} or higher.

Q.E.D.

Proof of Lemma 1.1:

Follows from equations (1), (3), and the equation for $\hat{\lambda}$, (insert number here).

Q.E.D.

Proof of Asymptotic Expansion 1.2:

(1) Follows from the Central Limit Theorem.

(2) Follows from the Central Limit Theorem.

(3) $\frac{1}{N}\sum_i (V_i-\tau)\hat{\lambda}'\psi_{2i}$

$$=\hat{\lambda}'\left(\frac{1}{N}\sum_i (V_i-\tau)\psi_{2i}\right)$$

$$=\hat{\lambda}'\sigma_{V\psi_2}+\hat{\lambda}'\left(\frac{1}{N}\sum_i\left((V_i-\tau)\psi_{2i}-\sigma_{V\psi_2}\right)\right)$$

$$= \hat{\lambda}'\sigma_{V\psi_2} + \left(\Omega^{-1}\frac{1}{N}\sum_k \psi_{2k}\right)'\left(\frac{1}{N}\sum_r\left((V_r - \tau)\psi_{2r} - \sigma_{V\psi_2}\right)\right),$$

$$-\left(\Omega^{-1}\mathrm{H}\Omega^{-1}\left(\frac{1}{N}\sum_s \psi_{2s}\right)\right)'\left(\frac{1}{N}\sum_i\left((V_i - \tau)\psi_{2i} - \sigma_{V\psi_2}\right)\right) + o_p\left(N^{-3/2}\right)$$

$$= \hat{\lambda}'\sigma_{V\psi_2} + R_{N\tau} + S_{N\tau} + o_p\left(n^{-3/2}\right).$$

where $R_{N\tau} = \left(\Omega^{-1}\frac{1}{N}\sum_s \psi_{2s}\right)'\left(\frac{1}{N}\sum_i\left((V_i - \tau)\psi_{2i} - \sigma_{V\psi_2}\right)\right) = O_p\left(N^{-1}\right)$ and

$S_{N\tau} = -\left(\Omega^{-1}\mathrm{H}\Omega^{-1}\left(\frac{1}{N}\sum_s \psi_{2s}\right)\right)'\left(\frac{1}{N}\sum_i\left((V_i - \tau)\psi_{2i} - \sigma_{V\psi_2}\right)\right) = O_p\left(N^{-3/2}\right)$. This is because the highest order of $\hat{\lambda}$ is $N^{-1/2}$, and (4) and (10) are $O_p\left(N^{-1/2}\right)$. Thus, the highest order of the product is N^{-1}, the next highest order is $N^{-3/2}$, and the other terms are $o_p\left(N^{-3/2}\right)$ or less.

(4) $\left(\frac{1}{N}\sum_i\left(1 - \hat{\lambda}'\psi_{2i}\right)\right)^{-1}$

$$= \left(1 - \hat{\lambda}'\left(\frac{1}{N}\sum_i \psi_{2i}\right)\right)^{-1}$$

$$= \left(1 + \hat{\lambda}'\left(\frac{1}{N}\sum_r \psi_{2r}\right)\right)\left(1 + \hat{\lambda}'\left(\frac{1}{N}\sum_s \psi_{2s}\right)\right)^{-1}\left(1 - \hat{\lambda}'\left(\frac{1}{N}\sum_i \psi_{2i}\right)\right)^{-1}$$

$$= \left(1 + \hat{\lambda}'\left(\frac{1}{N}\sum_r \psi_{2r}\right)\right)\left(\left(1 - \hat{\lambda}'\left(\frac{1}{N}\sum_s \psi_{2s}\right)\right)\left(1 + \hat{\lambda}'\left(\frac{1}{N}\sum_i \psi_{2i}\right)\right)\right)^{-1}$$

$$=\left(1+\hat{\lambda}'\left(\frac{1}{N}\sum_{s}\psi_{2s}\right)\right)\left(1-\left(\hat{\lambda}'\left(\frac{1}{N}\sum_{i}\psi_{2i}\right)\right)^{2}\right)^{-1}$$

$$=1+\hat{\lambda}'\left(\frac{1}{N}\sum_{i}\psi_{2i}\right)+o_p\left(N^{-3/2}\right).$$

Q.E.D.

PROOF OF THEOREM 1.1:

$\hat{\tau}-\tau$

$$=\left(\frac{1}{N}\sum_{r}(V_r-\tau)-\frac{1}{N}\sum_{s}(V_s-\tau)\hat{\lambda}'\psi_{2r}\right)\left(1+\hat{\lambda}'\left(\frac{1}{N}\sum_{i}\psi_{2i}\right)+o_p\left(N^{-3/2}\right)\right)$$

$$=\left(\left(\frac{1}{N}\sum_{r}(V_r-\tau)\right)-\left(\hat{\lambda}'\sigma_{V\psi_2}+R_{N\tau}+S_{N\tau}\right)\right)\left(1+\left(\Omega^{-1}\left(\frac{1}{N}\sum_{s}\psi_{2s}\right)\right)'\left(\frac{1}{N}\sum_{i}\psi_{2i}\right)\right)$$

$$=\left(\left(\frac{1}{N}\sum_{r}(V_r-\tau)\right)-\left((T_{N\lambda}+R_{N\lambda}+S_{N\lambda})'\sigma_{V\psi_2}+R_{N\tau}+S_{N\tau}\right)\right)\left(1+\left(\Omega^{-1}\left(\frac{1}{N}\sum_{s}\psi_{2s}\right)\right)'\left(\frac{1}{N}\sum_{i}\psi_{2i}\right)\right)$$

$$=T_{N1}+T_{N2}+R_{N1}+R_{N2}+S_{N1}+S_{N2}+S_{N3}+S_{N4}+o_p\left(N^{-3/2}\right).$$

The first equality holds because of the expansion of (insert number). The second equality arises from equation (insert number) and substitution of the order $N^{-1/2}$ term of $\hat{\lambda}$, $T_{N\lambda}$. The third equality substitutes equation (insert number), the equation for $\hat{\lambda}$ in the first factor. Finally, the fourth equality results from multiplication of the two factors and ignores terms that are of order less than $N^{-3/2}$.

Q.E.D.

Proof of Corollary 1.1:

T_{N1}

$$= \frac{1}{N}\sum_i (V_i - \tau),$$

T_{N2}

$$= -\sigma_{V\psi_2}{}'\Omega^{-1}\left(\frac{1}{N}\sum_i \psi_{2i}\right) = -\frac{1}{N}\sigma_{V\psi_2}{}'\Omega^{-1}\left(\sum_{i=1}^{N}\begin{pmatrix}\psi_{2i1}\\ \vdots \\ \psi_{2i,M-1}\end{pmatrix}\right),$$

R_{N1}

$$= \left(\frac{1}{N}\sum_i \psi_{2i}\right)'\Omega^{-1}\mathrm{H}\Omega^{-1}\sigma_{V\psi_2},$$

$$= \frac{1}{N^2}\left(\sum_{s=1}^{N}\begin{pmatrix}\psi_{2s1}\\ \vdots \\ \psi_{2s,M-1}\end{pmatrix}'\right)$$

$$\times\begin{pmatrix}\sum_{j=1}^{M-1}\sum_{i=1}^{N}\omega_{1j}{}^{-1}\left(\psi_{2ij}\psi_{2i1}-\omega_{1j}\right) & \cdots & \sum_{j=1}^{M-1}\sum_{i=1}^{N}\omega_{1j}{}^{-1}\left(\psi_{2ij}\psi_{2i,M-1}-\omega_{M-1,j}\right)\\ \vdots & \ddots & \vdots \\ \sum_{j=1}^{M-1}\sum_{i=1}^{N}\omega_{M-1,j}{}^{-1}\left(\psi_{2ij}\psi_{2i1}-\omega_{1j}\right) & \cdots & \sum_{j=1}^{M-1}\sum_{i=1}^{N}\omega_{M-1,j}{}^{-1}\left(\psi_{2ij}\psi_{2i,M-1}-\omega_{M-1,j}\right)\end{pmatrix}\Omega^{-1}\sigma_{V\psi_2}$$

$$=\frac{1}{N^2}\begin{pmatrix} \sum_{t=1}^{M-1}\sum_{j=1}^{M-1}\sum_{s=1}^{N}\sum_{i=1}^{N}\omega_{tj}^{-1}\psi_{2st}\left(\psi_{2i1}\psi_{2ij}-\omega_{1j}\right) \\ \vdots \\ \sum_{t=1}^{M-1}\sum_{j=1}^{M-1}\sum_{s=1}^{N}\sum_{i=1}^{N}\omega_{tj}^{-1}\psi_{2st}\left(\psi_{2i,M-1}\psi_{2ij}-\omega_{M-1,j}\right) \end{pmatrix}'\Omega^{-1}\sigma_{V\psi_2}$$

R_{N2}

$$=-\left(\frac{1}{N}\sum_{s}\psi_{2s}\right)'\Omega^{-1}\left(\frac{1}{N}\sum_{i}\left((V_i-\tau)\psi_{2i}-\sigma_{V\psi_2}\right)\right).$$

$$=-\frac{1}{N^2}\left(\sum_{s=1}^{N}\begin{pmatrix}\psi_{2s1}\\ \vdots \\ \psi_{2s,m-1}\end{pmatrix}'\right)\begin{pmatrix} \sum_{j=1}^{M-1}\sum_{i=1}^{N}\omega_{1j}^{-1}\left((V_i-\tau)\psi_{2ij}-\sigma_{V\psi_2 j}\right) \\ \vdots \\ \sum_{j=1}^{M-1}\sum_{i=1}^{N}\omega_{M-1,j}^{-1}\left((V_i-\tau)\psi_{2ij}-\sigma_{V\psi_2 j}\right) \end{pmatrix}$$

$$=-\frac{1}{N^2}\sum_{t=1}^{M-1}\sum_{j=1}^{M-1}\sum_{s=1}^{N}\sum_{i=1}^{N}\omega_{tj}^{-1}\psi_{2st}\left((V_i-\tau)\psi_{2ij}-\sigma_{V\psi_2 j}\right)$$

S_{N1}

$$=-\left(\frac{1}{N}\sum_{i}\psi_{2i}\right)'\Omega^{-1}\mathrm{H}\Omega^{-1}\mathrm{H}\Omega^{-1}\sigma_{V\psi_2}$$

$$=-\left(\frac{1}{N}\sum_{r}\psi_{2r}\right)'\Omega^{-1}\left(\frac{1}{N}\sum_{s}\left(\psi_{2s}\psi_{2s}'-\Omega\right)\right)\Omega^{-1}\left(\frac{1}{N}\sum_{i}\left(\psi_{2i}\psi_{2i}'-\Omega\right)\right)\Omega^{-1}\sigma_{V\psi_2}$$

$$= -\frac{1}{N^4}\begin{pmatrix} \sum_{p=1}^{M-1}\sum_{l=1}^{M-1}\sum_{r=1}^{N}\sum_{s=1}^{N}\psi_{2rp}\omega_{pl}^{-1}\left(\psi_{2s1}\psi_{2sl}-\omega_{1l}\right) \\ \vdots \\ \sum_{p=1}^{M-1}\sum_{l=1}^{M-1}\sum_{r=1}^{N}\sum_{s=1}^{N}\psi_{2rp}\omega_{pl}^{-1}\left(\psi_{2s,M-1}\psi_{2sl}-\omega_{M-1,l}\right) \end{pmatrix}'$$

$$\times\begin{pmatrix} \sum_{j=1}^{M-1}\sum_{i=1}^{N}\omega_{1j}^{-1}\left(\psi_{2ij}\psi_{2i1}-\omega_{1j}\right) & \cdots & \sum_{j=1}^{M-1}\sum_{i=1}^{N}\omega_{1j}^{-1}\left(\psi_{2ij}\psi_{2i,M-1}-\omega_{M-1,j}\right) \\ \vdots & \ddots & \vdots \\ \sum_{j=1}^{M-1}\sum_{i=1}^{N}\omega_{M-1,j}^{-1}\left(\psi_{2ij}\psi_{2i1}-\omega_{1j}\right) & \cdots & \sum_{j=1}^{M-1}\sum_{i=1}^{N}\omega_{M-1,j}^{-1}\left(\psi_{2ij}\psi_{2i,M-1}-\omega_{M-1,j}\right) \end{pmatrix}\Bigg)\Omega^{-1}\sigma_{V\psi_2}$$

$$= -\frac{1}{N^4}\begin{pmatrix} \sum_{p=1}^{M-1}\sum_{l=1}^{M-1}\sum_{t=1}^{M-1}\sum_{j=1}^{M-1}\sum_{r=1}^{N}\sum_{s=1}^{N}\sum_{i=1}^{N}\psi_{2rp}\omega_{pl}^{-1}\left(\psi_{2st}\psi_{2sl}-\omega_{tl}\right)\omega_{tj}^{-1}\left(\psi_{2ij}\psi_{2i1}-\omega_{1j}\right) \\ \vdots \\ \sum_{p=1}^{M-1}\sum_{l=1}^{M-1}\sum_{t=1}^{M-1}\sum_{j=1}^{M-1}\sum_{r=1}^{N}\sum_{s=1}^{N}\sum_{i=1}^{N}\psi_{2rp}\omega_{pl}^{-1}\left(\psi_{2st}\psi_{2sl}-\omega_{tl}\right)\omega_{tj}^{-1}\left(\psi_{2ij}\psi_{2i,M-1}-\omega_{M-1,j}\right) \end{pmatrix}'\Omega^{-1}\sigma_{V\psi_2}$$

S_{N2}

$$=\left(\frac{1}{N}\sum_{s}\psi_{2s}\right)'\Omega^{-1}H\Omega^{-1}\left(\frac{1}{N}\sum_{i}\left(\left(V_i-\tau\right)\psi_{2i}-\sigma_{V\psi_2}\right)\right)$$

$$=\left(\frac{1}{N}\sum_{r}\psi_{2r}\right)'\Omega^{-1}\left(\frac{1}{N}\sum_{s}\left(\psi_{2s}\psi_{2s}'-\Omega\right)\right)\Omega^{-1}\left(\frac{1}{N}\sum_{i}\left(\left(V_i-\tau\right)\psi_{2i}-\sigma_{V\psi_2}\right)\right)$$

$$=\frac{1}{N^3}\begin{pmatrix} \sum_{p=1}^{M-1}\sum_{l=1}^{M-1}\sum_{r=1}^{N}\sum_{s=1}^{N}\psi_{2rp}\omega_{pl}^{-1}\left(\psi_{2s1}\psi_{2sl}-\omega_{1l}\right) \\ \vdots \\ \sum_{p=1}^{M-1}\sum_{l=1}^{M-1}\sum_{r=1}^{N}\sum_{s=1}^{N}\psi_{2rp}\omega_{pl}^{-1}\left(\psi_{2s,M-1}\psi_{2sl}-\omega_{M-1,l}\right) \end{pmatrix}'\begin{pmatrix} \sum_{j=1}^{M-1}\sum_{i=1}^{N}\omega_{1j}^{-1}\left(\left(V_i-\tau\right)\psi_{2ij}-\sigma_{V\psi_2 j}\right) \\ \vdots \\ \sum_{j=1}^{M-1}\sum_{i=1}^{N}\omega_{M-1,j}^{-1}\left(\left(V_i-\tau\right)\psi_{2ij}-\sigma_{V\psi_2 j}\right) \end{pmatrix}$$

$$=\frac{1}{N^3}\left(\sum_{p=1}^{M-1}\sum_{l=1}^{M-1}\sum_{t=1}^{M-1}\sum_{j=1}^{M-1}\sum_{r=1}^{N}\sum_{s=1}^{N}\sum_{i=1}^{N}\psi_{2rp}{\omega_{pl}}^{-1}\left(\psi_{2st}\psi_{2sl}-\omega_{tl}\right){\omega_{tj}}^{-1}\left(\left(V_i-\tau\right)\psi_{2ij}-\sigma_{V\psi_2 j}\right)\right)$$

S_{N3}

$$=\left(\frac{1}{N}\sum_r\left(V_r-\tau\right)\right)\left(\frac{1}{N}\sum_s\psi_{2s}\right)'\Omega^{-1}\left(\frac{1}{N}\sum_i\psi_{2i}\right)$$

$$=\frac{1}{N^3}\left(\sum_{r=1}^{N}\left(V_r-\tau\right)\right)\left(\sum_{t=1}^{M-1}\sum_{s=1}^{N}\sum_{j=1}^{M-1}\sum_{i=1}^{N}\psi_{2st}{\omega_{tj}}^{-1}\psi_{2ij}\right)$$

$$=\frac{1}{N^3}\sum_{t=1}^{M-1}\sum_{j=1}^{M-1}\sum_{r=1}^{N}\sum_{s=1}^{N}\sum_{i=1}^{N}{\omega_{tj}}^{-1}\left(V_r-\tau\right)\psi_{2st}\psi_{2ij}$$

S_{N4}

$$=-{\sigma_{V\psi_2}}'\Omega^{-1}\left(\frac{1}{N}\sum_r\psi_{2r}\right)\left(\frac{1}{N}\sum_s\psi_{2s}\right)'\Omega^{-1}\left(\frac{1}{N}\sum_i\psi_{2i}\right)$$

$$=-\frac{1}{N^3}\left(\sum_{p=1}^{M-1}\sum_{l=1}^{M-1}\sum_{r=1}^{N}\sigma_{V\psi_2,p}{\omega_{pl}}^{-1}\psi_{2rl}\right)\left(\sum_{t=1}^{M-1}\sum_{j=1}^{M-1}\sum_{s=1}^{M}\sum_{i=1}^{N}\psi_{2st}{\omega_{tj}}^{-1}\psi_{2ij}\right)$$

$$=-\frac{1}{N^3}\left(\sum_{p=1}^{M-1}\sum_{l=1}^{M-1}\sum_{t=1}^{M-1}\sum_{j=1}^{M-1}\sum_{r=1}^{N}\sum_{s=1}^{N}\sum_{i=1}^{N}\sigma_{V\psi_2,p}{\omega_{pl}}^{-1}\psi_{2rl}\psi_{2st}{\omega_{tj}}^{-1}\psi_{2ij}\right)$$

$E\left(\left(T_{N1}\right)^2\right)$

$$=E\left(\left(\frac{1}{N}\sum_i\left(V_i-\tau\right)\right)^2\right)$$

$$=\frac{1}{N^2}E\left(\sum_{s=1}^{N}\sum_{t=1}^{N}(V_s-\tau)(V_t-\tau)\right)$$

$$=\frac{1}{N^2}E\left(\sum_{t=1}^{N}(V_t-\tau)^2\right)$$

$$=\frac{\sigma_V^{\,2}}{N}.$$

$$E\left((T_{N2})^2\right)$$

$$=E\left(\left(-\sigma_{V\psi_2}'\Omega^{-1}\left(\frac{1}{N}\sum_i\psi_{2i}\right)\right)^2\right)$$

$$=\frac{1}{N^2}E\left(\sigma_{V\psi_2}'\Omega^{-1}\left(\sum_{s=1}^{N}\psi_{2s}\right)\left(\sum_{t=1}^{N}\psi_{2t}\right)'\Omega^{-1}\sigma_{V\psi_2}\right)$$

$$=\frac{1}{N^2}\left(\sigma_{V\psi_2}'\Omega^{-1}E\left(\sum_{s=1}^{N}\sum_{t=1}^{N}\psi_{2s}\psi_{2t}'\right)\Omega^{-1}\sigma_{V\psi_2}\right)$$

$$=\frac{1}{N^2}\left(\sigma_{V\psi_2}'\Omega^{-1}E\left(\sum_{t=1}^{N}\psi_{2t}\psi_{2t}'\right)\Omega^{-1}\sigma_{V\psi_2}\right)$$

$$=\frac{1}{N}\sigma_{V\psi_2}'\Omega^{-1}\sigma_{V\psi_2}.$$

$$E(T_{N1}T_{N2})$$

$$=-E\left(\left(\frac{1}{N}\sum_s(V_s-\tau)\right)\sigma_{V\psi_2}'\Omega^{-1}\left(\frac{1}{N}\sum_t\psi_{2t}\right)\right)$$

$$=-\frac{1}{N^2}E\left(\left(\sum_{s=1}^{N}(V_s-\tau)\right)\left(\sum_{t=1}^{N}\psi_{2t}\right)'\Omega^{-1}\sigma_{V\psi_2}\right)$$

$$=-\frac{1}{N^2}\left(\sum_{s=1}^{N}\sum_{i=1}^{N}E\left((V_s-\tau)\psi_{2i}'\right)\right)\Omega^{-1}\sigma_{V\psi_2}$$

$$=-\frac{1}{N^2}\left(\sum_{i=1}^{N}E\left((V_i-\tau)\psi_{2i}'\right)\right)\Omega^{-1}\sigma_{V\psi_2}$$

$$=-\frac{1}{N}\sigma_{V\psi_2}'\Omega^{-1}\sigma_{V\psi_2}$$

$E(T_{N1}R_{N1})$

$$=E\left(\left(\frac{1}{N}\sum_{r}(V_r-\tau)\right)\frac{1}{N^2}\begin{pmatrix}\sum_{t=1}^{M-1}\sum_{j=1}^{M-1}\sum_{s=1}^{N}\sum_{i=1}^{N}\omega_{tj}^{-1}\psi_{2st}\left(\psi_{2i1}\psi_{2ij}-\omega_{1j}\right)\\ \vdots \\ \sum_{t=1}^{M-1}\sum_{j=1}^{M-1}\sum_{s=1}^{N}\sum_{i=1}^{N}\omega_{tj}^{-1}\psi_{2st}\left(\psi_{2i,M-1}\psi_{2ij}-\omega_{M-1,j}\right)\end{pmatrix}'\Omega^{-1}\sigma_{V\psi_2}\right)$$

$$=\frac{1}{N^3}E\begin{pmatrix}\sum_{t=1}^{M-1}\sum_{j=1}^{M-1}\sum_{r=1}^{N}\sum_{s=1}^{N}\sum_{i=1}^{N}\omega_{tj}^{-1}(V_r-\tau)\psi_{2st}\left(\psi_{2i1}\psi_{2ij}-\omega_{1j}\right)\\ \vdots \\ \sum_{t=1}^{M-1}\sum_{j=1}^{M-1}\sum_{r=1}^{N}\sum_{s=1}^{N}\sum_{i=1}^{N}\omega_{tj}^{-1}(V_r-\tau)\psi_{2st}\left(\psi_{2i,M-1}\psi_{2ij}-\omega_{M-1,j}\right)\end{pmatrix}'\Omega^{-1}\sigma_{V\psi_2}$$

Remember that $E(V_i-\tau)=0, E(\psi_{2i})=0$, and $\mathrm{E}\left(\psi_{2i}\psi_{2i}'-\Omega\right)=0$. Therefore, the expectation of the individual sums equal zero whenever $r\neq s$, $s\neq i$, or $r\neq i$. Thus, there is one case that has a non-zero expectation, $r=s=i$. Hence,

$$=\frac{1}{N^3}E\begin{pmatrix}\sum_{t=1}^{M-1}\sum_{j=1}^{M-1}\sum_{i=1}^{N}\omega_{tj}^{-1}(V_i-\tau)\psi_{2it}\left(\psi_{2i1}\psi_{2ij}-\omega_{1j}\right)\\ \vdots\\ \sum_{t=1}^{M-1}\sum_{j=1}^{M-1}\sum_{i=1}^{N}\omega_{tj}^{-1}(V_i-\tau)\psi_{2it}\left(\psi_{2i,M-1}\psi_{2ij}-\omega_{M-1,j}\right)\end{pmatrix}'\Omega^{-1}\sigma_{V\psi_2}$$

$$=\frac{1}{N^2}\begin{pmatrix}\sum_{t=1}^{M-1}\sum_{j=1}^{M-1}\omega_{tj}^{-1}Cov\left(\left(V_i-\tau\right)\psi_{2it},\psi_{2i1}\psi_{2ij}\right)\\ \vdots\\ \sum_{t=1}^{M-1}\sum_{j=1}^{M-1}\omega_{tj}^{-1}Cov\left(\left(V_i-\tau\right)\psi_{2it},\psi_{2i,M-1}\psi_{2ij}\right)\end{pmatrix}'\Omega^{-1}\sigma_{V\psi_2}$$

$$E\left(T_{N2}R_{N1}\right)$$

$$=-\frac{1}{N}\sigma_{V\psi_2}'\Omega^{-1}E\left(\left(\sum_{r=1}^{N}\begin{pmatrix}\psi_{2r1}\\ \vdots\\ \psi_{2r,M-1}\end{pmatrix}\right)\cdot\frac{1}{N^2}\begin{pmatrix}\sum_{t=1}^{M-1}\sum_{j=1}^{M-1}\sum_{s=1}^{N}\sum_{i=1}^{N}\omega_{tj}^{-1}\psi_{2st}\left(\psi_{2i1}\psi_{2ij}-\omega_{1j}\right)\\ \vdots\\ \sum_{t=1}^{M-1}\sum_{j=1}^{M-1}\sum_{s=1}^{N}\sum_{i=1}^{N}\omega_{tj}^{-1}\psi_{2st}\left(\psi_{2i,M-1}\psi_{2ij}-\omega_{M-1,j}\right)\end{pmatrix}'\right)\Omega^{-1}\sigma_{V\psi_2}$$

$$=-\frac{1}{N^3}\sigma_{V\psi_2}'\Omega^{-1}$$

$$\times E\begin{pmatrix}\sum_{t=1}^{M-1}\sum_{j=1}^{M-1}\sum_{r=1}^{N}\sum_{s=1}^{N}\sum_{i=1}^{N}\omega_{tj}^{-1}\psi_{2r1}\psi_{2st}\left(\psi_{2i1}\psi_{2ij}-\omega_{1j}\right) & \cdots & \sum_{t=1}^{M-1}\sum_{j=1}^{M-1}\sum_{r=1}^{N}\sum_{s=1}^{N}\sum_{i=1}^{N}\omega_{tj}^{-1}\psi_{2r1}\psi_{2st}\big(\psi_{2i,M-1}\psi_{2ij}-\omega_{M}\\ \vdots & \ddots & \vdots\\ \sum_{t=1}^{M-1}\sum_{j=1}^{M-1}\sum_{r=1}^{N}\sum_{s=1}^{N}\sum_{i=1}^{N}\omega_{tj}^{-1}\psi_{2r,M-1}\psi_{2st}\left(\psi_{2i1}\psi_{2ij}-\omega_{1j}\right) & \cdots & \sum_{t=1}^{M-1}\sum_{j=1}^{M-1}\sum_{r=1}^{N}\sum_{s=1}^{N}\sum_{i=1}^{N}\omega_{tj}^{-1}\psi_{2r,M-1}\psi_{2st}\big(\psi_{2i,M-1}\psi_{2ij}-\omega\end{pmatrix}$$

$$\times\Omega^{-1}\sigma_{V\psi_2}$$

$$
\begin{aligned}
&= -\frac{1}{N^3}\sigma_{V\psi_2}{}'\Omega^{-1} \\
&\times E\begin{pmatrix}
\sum_{t=1}^{M-1}\sum_{j=1}^{M-1}\sum_{i=1}^{N}\omega_{tj}{}^{-1}\psi_{2i1}\psi_{2it}\left(\psi_{2i1}\psi_{2ij}-\omega_{1j}\right) & \cdots & \sum_{t=1}^{M-1}\sum_{j=1}^{M-1}\sum_{i=1}^{N}\omega_{tj}{}^{-1}\psi_{2i1}\psi_{2it}\left(\psi_{2i,M-1}\psi_{2ij}-\omega_{M-1,j}\right) \\
\vdots & \ddots & \vdots \\
\sum_{t=1}^{M-1}\sum_{j=1}^{M-1}\sum_{i=1}^{N}\omega_{tj}{}^{-1}\psi_{2i,M-1}\psi_{2it}\left(\psi_{2i1}\psi_{2ij}-\omega_{1j}\right) & \cdots & \sum_{t=1}^{M-1}\sum_{j=1}^{M-1}\sum_{i=1}^{N}\omega_{tj}{}^{-1}\psi_{2i,M-1}\psi_{2it}\left(\psi_{2i,M-1}\psi_{2ij}-\omega_{M-1,j}\right)
\end{pmatrix} \\
&\times \Omega^{-1}\sigma_{V\psi_2}
\end{aligned}
$$

$$
= -\frac{1}{N^2}\sigma_{V\psi_2}{}'\Omega^{-1}\begin{pmatrix}
\sum_{t=1}^{M-1}\sum_{j=1}^{M-1}\omega_{tj}{}^{-1}Cov\left(\psi_{2i1}\psi_{2it},\psi_{2i1}\psi_{2ij}\right) & \cdots & \sum_{t=1}^{M-1}\sum_{j=1}^{M-1}\omega_{tj}{}^{-1}Cov\left(\psi_{2i1}\psi_{2it},\psi_{2i,M-1}\psi_{2ij}\right) \\
\vdots & \ddots & \vdots \\
\sum_{t=1}^{M-1}\sum_{j=1}^{M-1}\omega_{tj}{}^{-1}Cov\left(\psi_{2i,M-1}\psi_{2it},\psi_{2i1}\psi_{2ij}\right) & \cdots & \sum_{t=1}^{M-1}\sum_{j=1}^{M-1}\omega_{tj}{}^{-1}Cov\left(\psi_{2i,M-1}\psi_{2it},\psi_{2i,M-1}\psi_{2ij}\right)
\end{pmatrix}\Omega^{-}
$$

$$
\begin{aligned}
&E(T_{N1}R_{N2}) \\
&= -E\left(\frac{1}{N}\left(\sum_{r}(V_r-\tau)\right)\frac{1}{N^2}\left(\sum_{t=1}^{M-1}\sum_{j=1}^{M-1}\sum_{s=1}^{N}\sum_{i=1}^{N}\omega_{tj}{}^{-1}\psi_{2st}\left(\left(V_i-\tau\right)\psi_{2ij}-\sigma_{V\psi_2j}\right)\right)\right) \\
&= -\frac{1}{N^3}E\left(\sum_{t=1}^{M-1}\sum_{j=1}^{M-1}\sum_{r=1}^{N}\sum_{s=1}^{N}\sum_{i=1}^{N}\omega_{tj}{}^{-1}\left(V_r-\tau\right)\psi_{2st}\left(\left(V_i-\tau\right)\psi_{2ij}-\sigma_{V\psi_2j}\right)\right) \\
&= -\frac{1}{N^3}E\left(\sum_{t=1}^{M-1}\sum_{j=1}^{M-1}\sum_{i=1}^{N}\omega_{tj}{}^{-1}\left(V_i-\tau\right)\psi_{2it}\left(\left(V_i-\tau\right)\psi_{2ij}-\sigma_{V\psi_2j}\right)\right) \\
&= -\frac{1}{N^2}\left(\sum_{t=1}^{M-1}\sum_{j=1}^{M-1}\omega_{tj}{}^{-1}Cov\left(\left(V_i-\tau\right)\psi_{2it},\left(V_i-\tau\right)\psi_{2ij}\right)\right)
\end{aligned}
$$

$E(T_{N2}R_{N2})$

$$=\frac{1}{N}\sigma_{V\psi_2}{}'\Omega^{-1}E\left(\sum_{r=1}^{N}\begin{pmatrix}\psi_{2r1}\\ \vdots\\ \psi_{2r,M-1}\end{pmatrix}\cdot\frac{1}{N^2}\left(\sum_{t=1}^{M-1}\sum_{j=1}^{M-1}\sum_{s=1}^{N}\sum_{t=1}^{N}\omega_{ij}{}^{-1}\psi_{2st}\left(\left(V_i-\tau\right)\psi_{2ij}-\sigma_{V\psi_2j}\right)\right)\right)$$

$$=\frac{1}{N^3}\sigma_{V\psi_2}{}'\Omega^{-1}E\begin{pmatrix}\sum_{t=1}^{M-1}\sum_{j=1}^{M-1}\sum_{r=1}^{N}\sum_{s=1}^{N}\sum_{i=1}^{N}\psi_{2r1}\psi_{2st}\omega_{tj}{}^{-1}\left(\left(V_i-\tau\right)\psi_{2ij}-\sigma_{V\psi_2j}\right)\\ \vdots\\ \sum_{t=1}^{M-1}\sum_{j=1}^{M-1}\sum_{r=1}^{N}\sum_{s=1}^{N}\sum_{i=1}^{N}\psi_{2r,M-1}\psi_{2st}\omega_{tj}{}^{-1}\left(\left(V_i-\tau\right)\psi_{2ij}-\sigma_{V\psi_2j}\right)\end{pmatrix}$$

$$=\frac{1}{N^3}\sigma_{V\psi_2}{}'\Omega^{-1}E\begin{pmatrix}\sum_{t=1}^{M-1}\sum_{j=1}^{M-1}\sum_{i=1}^{N}\omega_{tj}{}^{-1}\psi_{2i1}\psi_{2it}\left(\left(V_i-\tau\right)\psi_{2ij}-\sigma_{V\psi_2j}\right)\\ \vdots\\ \sum_{t=1}^{M-1}\sum_{j=1}^{M-1}\sum_{i=1}^{N}\omega_{tj}{}^{-1}\psi_{2i,M-1}\psi_{2it}\left(\left(V_i-\tau\right)\psi_{2ij}-\sigma_{V\psi_2j}\right)\end{pmatrix}$$

$$=\frac{1}{N^2}\sigma_{V\psi_2}{}'\Omega^{-1}\begin{pmatrix}\sum_{t=1}^{M-1}\sum_{j=1}^{M-1}\omega_{tj}{}^{-1}Cov\left(\psi_{2i1}\psi_{2it},\left(V_i-\tau\right)\psi_{2ij}\right)\\ \vdots\\ \sum_{t=1}^{M-1}\sum_{j=1}^{M-1}\omega_{tj}{}^{-1}Cov\left(\psi_{2i,M-1}\psi_{2it},\left(V_i-\tau\right)\psi_{2ij}\right)\end{pmatrix}$$

$=E\left(T_{N1}R_{N1}\right)$

$E\left(R_{N1}R_{N2}\right)$

$$= -\frac{1}{N^2}\sigma_{V\psi_2}{}'\Omega^{-1}E\left(\begin{pmatrix} \sum_{p=1}^{M-1}\sum_{l=1}^{M-1}\sum_{k=1}^{N}\sum_{r=1}^{N}\omega_{pl}{}^{-1}\psi_{2kp}\left(\psi_{2r1}\psi_{2rl}-\omega_{1l}\right) \\ \vdots \\ \sum_{p=1}^{M-1}\sum_{l=1}^{M-1}\sum_{k=1}^{N}\sum_{r=1}^{N}\omega_{pl}{}^{-1}\psi_{2kp}\left(\psi_{2r,M-}\psi_{2rl}-\omega_{M-1,l}\right) \end{pmatrix}\right.$$

$$\left.\times\frac{1}{N^2}\left(\sum_{t=1}^{M-1}\sum_{j=1}^{M-1}\sum_{s=1}^{N}\sum_{i=1}^{N}\omega_{tj}{}^{-1}\psi_{2st}\left(\left(V_i-\tau\right)\psi_{2ij}-\sigma_{V\psi_2j}\right)\right)\right)$$

$$= -\frac{1}{N^4}\sigma_{V\psi_2}{}'\Omega^{-1}E\left(\begin{matrix} \sum_{p=1}^{M-1}\sum_{l=1}^{M-1}\sum_{t=1}^{M-1}\sum_{j=1}^{M-1}\sum_{k=1}^{N}\sum_{r=1}^{N}\sum_{s=1}^{N}\sum_{i=1}^{N}\omega_{pl}{}^{-1}\omega_{tj}{}^{-1}\psi_{2kp}\psi_{2st}\left(\psi_{2r1}\psi_{2rl}-\omega_{1l}\right)\left(\left(V_i-\tau\right)\psi_{2ij}-\sigma_{V\psi_2j}\right) \\ \vdots \\ \sum_{p=1}^{M-1}\sum_{l=1}^{M-1}\sum_{t=1}^{M-1}\sum_{j=1}^{M-1}\sum_{k=1}^{N}\sum_{r=1}^{N}\sum_{s=1}^{N}\sum_{i=1}^{N}\omega_{pl}{}^{-1}\omega_{tj}{}^{-1}\psi_{2kp}\psi_{2st}\left(\psi_{2r,M-1}\psi_{2rl}-\omega_{M-1,l}\right)\left(\left(V_i-\tau\right)\psi_{2ij}-\sigma_{V\psi_2j}\right. \end{matrix}\right.$$

This expectation equals zero whenever one of the four indices, k, r, s, and i, does not equal one of the others. Also, we can ignore the case when $k = r = s = i$ because it is of order N^{-3}. Thus, we consider the three cases when $(k = r, s = i, r \neq i)$, $(k = i, r = s, s \neq i)$, and $(k = s, r = i, s \neq i)$. The equation above becomes

$$= -\frac{1}{N^4}\sigma_{V\psi_2}{}'\Omega^{-1}E\begin{pmatrix} \sum_{p=1}^{M-1}\sum_{l=1}^{M-1}\sum_{t=1}^{M-1}\sum_{j=1}^{M-1}\sum_{i=1}^{N}\sum_{r\neq i}\omega_{pl}{}^{-1}\omega_{tj}{}^{-1}\psi_{2rp}\left(\psi_{2r1}\psi_{2rl}-\omega_{1l}\right)\psi_{2it}\left(\left(V_i-\tau\right)\psi_{2ij}-\sigma_{V\psi_2j}\right) \\ \vdots \\ \sum_{p=1}^{M-1}\sum_{l=1}^{M-1}\sum_{t=1}^{M-1}\sum_{j=1}^{M-1}\sum_{i=1}^{N}\sum_{r\neq i}\omega_{pl}{}^{-1}\omega_{tj}{}^{-1}\psi_{2rp}\left(\psi_{2r,M-1}\psi_{2rl}-\omega_{M-1,l}\right)\psi_{2it}\left(\left(V_i-\tau\right)\psi_{2ij}-\sigma_{V\psi_2j}\right) \end{pmatrix}$$

$$-\frac{1}{N^4}\sigma_{V\psi_2}{}'\Omega^{-1}E\begin{pmatrix} \sum_{p=1}^{M-1}\sum_{l=1}^{M-1}\sum_{t=1}^{M-1}\sum_{j=1}^{M-1}\sum_{i=1}^{N}\sum_{s\neq i}\omega_{pl}{}^{-1}\omega_{tj}{}^{-1}\psi_{2st}\left(\psi_{2s1}\psi_{2sl}-\omega_{1l}\right)\psi_{2ip}\left(\left(V_i-\tau\right)\psi_{2ij}-\sigma_{V\psi_2 j}\right) \\ \vdots \\ \sum_{p=1}^{M-1}\sum_{l=1}^{M-1}\sum_{t=1}^{M-1}\sum_{j=1}^{M-1}\sum_{i=1}^{N}\sum_{s\neq i}\omega_{pl}{}^{-1}\omega_{tj}{}^{-1}\psi_{2st}\left(\psi_{2s,M-1}\psi_{2sl}-\omega_{M-1,l}\right)\psi_{2ip}\left(\left(V_i-\tau\right)\psi_{2ij}-\sigma_{V\psi_2 j}\right) \end{pmatrix}$$

$$-\frac{1}{N^4}\sigma_{V\psi_2}{}'\Omega^{-1}E\begin{pmatrix} \sum_{p=1}^{M-1}\sum_{l=1}^{M-1}\sum_{t=1}^{M-1}\sum_{j=1}^{M-1}\sum_{i=1}^{N}\sum_{s\neq i}\omega_{pl}{}^{-1}\omega_{tj}{}^{-1}\psi_{2sp}\psi_{2st}\left(\psi_{2i1}\psi_{2il}-\omega_{1l}\right)\left(\left(V_i-\tau\right)\psi_{2ij}-\sigma_{V\psi_2 j}\right) \\ \vdots \\ \sum_{p=1}^{M-1}\sum_{l=1}^{M-1}\sum_{t=1}^{M-1}\sum_{j=1}^{M-1}\sum_{i=1}^{N}\sum_{s\neq i}\omega_{pl}{}^{-1}\omega_{tj}{}^{-1}\psi_{2sp}\psi_{2st}\left(\psi_{2i,M-1}\psi_{2il}-\omega_{M-1,l}\right)\left(\left(V_i-\tau\right)\psi_{2ij}-\sigma_{V\psi_2 j}\right) \end{pmatrix}$$

$$+o_p\left(N^{-2}\right)$$

$$=-\frac{1}{N^2}\sigma_{V\psi_2}{}'\Omega^{-1}\begin{pmatrix} \sum_{p=1}^{M-1}\sum_{l=1}^{M-1}\sum_{t=1}^{M-1}\sum_{j=1}^{M-1}\omega_{pl}{}^{-1}\omega_{tj}{}^{-1}Cov\left(\psi_{2ip},\psi_{2i1}\psi_{2il}\right)Cov\left(\psi_{2it},\left(V_i-\tau\right)\psi_{2ij}\right) \\ \vdots \\ \sum_{p=1}^{M-1}\sum_{l=1}^{M-1}\sum_{t=1}^{M-1}\sum_{j=1}^{M-1}\omega_{pl}{}^{-1}\omega_{tj}{}^{-1}Cov\left(\psi_{2ip},\psi_{2i,M-1}\psi_{2il}\right)Cov\left(\psi_{2it},\left(V_i-\tau\right)\psi_{2ij}\right) \end{pmatrix}$$

$$-\frac{1}{N^2}\sigma_{V\psi_2}{}'\Omega^{-1}\begin{pmatrix} \sum_{p=1}^{M-1}\sum_{l=1}^{M-1}\sum_{t=1}^{M-1}\sum_{j=1}^{M-1}\omega_{pl}{}^{-1}\omega_{tj}{}^{-1}Cov\left(\psi_{2it},\psi_{2i1}\psi_{2il}\right)Cov\left(\psi_{2ip},\left(V_i-\tau\right)\psi_{2ij}\right) \\ \vdots \\ \sum_{p=1}^{M-1}\sum_{l=1}^{M-1}\sum_{t=1}^{M-1}\sum_{j=1}^{M-1}\omega_{pl}{}^{-1}\omega_{tj}{}^{-1}Cov\left(\psi_{2it},\psi_{2i,M-1}\psi_{2il}\right)Cov\left(\psi_{2ip},\left(V_i-\tau\right)\psi_{2ij}\right) \end{pmatrix}$$

$$-\frac{1}{N^2}\sigma_{V\psi_2}{}'\Omega^{-1}\begin{pmatrix} \sum_{p=1}^{M-1}\sum_{l=1}^{M-1}\sum_{t=1}^{M-1}\sum_{j=1}^{M-1}\omega_{pl}{}^{-1}\omega_{tj}{}^{-1}Cov\left(\psi_{2ip},\psi_{2it}\right)Cov\left(\psi_{2i1}\psi_{2il},\left(V_i-\tau\right)\psi_{2ij}\right) \\ \vdots \\ \sum_{p=1}^{M-1}\sum_{l=1}^{M-1}\sum_{t=1}^{M-1}\sum_{j=1}^{M-1}\omega_{pl}{}^{-1}\omega_{tj}{}^{-1}Cov\left(\psi_{2ip},\psi_{2it}\right)Cov\left(\psi_{2i,M-1}\psi_{2il},\left(V_i-\tau\right)\psi_{2ij}\right) \end{pmatrix}$$

$$+o_p\left(N^{-2}\right)$$

$$E\left(R_{v1}^{\;2}\right)$$

$$= \frac{1}{N^2}\sigma_{V\psi_2}{}'\Omega^{-1}E\left(\begin{pmatrix} \sum_{p=1}^{M-1}\sum_{l=1}^{M-1}\sum_{k=1}^{N}\sum_{r=1}^{N}\omega_{pl}^{\;-1}\psi_{2kp}\left(\psi_{2r1}\psi_{2rl}-\omega_{1l}\right) \\ \vdots \\ \sum_{p=1}^{M-1}\sum_{l=1}^{M-1}\sum_{k=1}^{N}\sum_{r=1}^{N}\omega_{pl}^{\;-1}\psi_{2kp}\left(\psi_{2r,M-}\psi_{2rl}-\omega_{M-1,l}\right) \end{pmatrix}\right.$$

$$\left.\times\frac{1}{N^2}\begin{pmatrix} \sum_{t=1}^{M-1}\sum_{j=1}^{M-1}\sum_{s=1}^{N}\sum_{i=1}^{N}\omega_{tj}^{\;-1}\psi_{2st}\left(\psi_{2i1}\psi_{2ij}-\omega_{1j}\right) \\ \vdots \\ \sum_{t=1}^{M-1}\sum_{j=1}^{M-1}\sum_{s=1}^{N}\sum_{i=1}^{N}\omega_{tj}^{\;-1}\psi_{2st}\left(\psi_{2i,M-}\psi_{2ij}-\omega_{M-1,j}\right) \end{pmatrix}'\right)\Omega^{-1}\sigma_{V\psi_2}$$

$$= \frac{1}{N^4}\sigma_{V\psi_2}{}'\Omega^{-1}E\begin{pmatrix} \sum_{p=1}^{M-1}\sum_{l=1}^{M-1}\sum_{t=1}^{M-1}\sum_{j=1}^{M-1}\sum_{k=1}^{N}\sum_{r=1}^{N}\sum_{s=1}^{N}\sum_{i=1}^{N}\omega_{pl}^{\;-1}\omega_{tj}^{\;-1}\psi_{2kp}\psi_{2st}\left(\psi_{2r1}\psi_{2rl}-\omega_{1l}\right)\left(\psi_{2i1}\psi_{2ij}-\omega_{1j}\right) & \cdots & \sum_{p=1}^{M-1}\sum_{l=1}^{M-1}\sum_{t=1}^{M-1}\sum_{j=1}^{M-1}\sum_{k=1}^{N}\sum_{r=1}^{N}\sum_{s=1}^{N}\sum_{i=1}^{N}\omega_{pl}^{\;-1}\omega_{tj}^{\;-1}\psi_{2kp}\psi_{2st}\left(\psi_{2r1}\psi_{2rl}-\omega_{1l}\right)\left(\psi_{2i,M-}\psi_{2ij}-\omega_{M-1,j}\right) \\ \vdots & \ddots & \vdots \\ \sum_{p=1}^{M-1}\sum_{l=1}^{M-1}\sum_{t=1}^{M-1}\sum_{j=1}^{M-1}\sum_{k=1}^{N}\sum_{r=1}^{N}\sum_{s=1}^{N}\sum_{i=1}^{N}\omega_{pl}^{\;-1}\omega_{tj}^{\;-1}\psi_{2kp}\psi_{2st}\left(\psi_{2r,M-1}\psi_{2rl}-\omega_{M-1,l}\right)\left(\psi_{2i1}\psi_{2ij}-\omega_{1j}\right) & \cdots & \sum_{p=1}^{M-1}\sum_{l=1}^{M-1}\sum_{t=1}^{M-1}\sum_{j=1}^{M-1}\sum_{k=1}^{N}\sum_{r=1}^{N}\sum_{s=1}^{N}\sum_{i=1}^{N}\omega_{pl}^{\;-1}\omega_{tj}^{\;-1}\psi_{2kp}\psi_{2st}\left(\psi_{2r,M-1}\psi_{2rl}-\omega_{M-1,l}\right)\left(\psi_{2i,M}\psi_{2ij}-\omega_{M-1,j}\right) \end{pmatrix}\Omega^{-1}\sigma_{V\psi_2}$$

For the same reason as mentioned above, we consider the three cases when $(k=r, s=i, r\neq i)$, $(k=i, r=s, s\neq i)$, and $(k=s, r=i, s\neq i)$. The equation above becomes

$$=\frac{1}{N^4}\sigma_{V\psi_2}{}'\Omega^{-1}E\left(\begin{array}{ccc} \sum_{p=1}^{M-1}\sum_{l=1}^{M-1}\sum_{t=1}^{M-1}\sum_{j=1}^{M-1}\sum_{i=1}^{N}\sum_{r\neq i}\omega_{pl}{}^{-1}\omega_{tj}{}^{-1}\psi_{2rp}\left(\psi_{2r1}\psi_{2rl}-\omega_{1l}\right)\psi_{2it}\left(\psi_{2i1}\psi_{2ij}-\omega_{1j}\right) & \cdots & \sum_{p=1}^{M-1}\sum_{l=1}^{M-1}\sum_{t=1}^{M-1}\sum_{j=1}^{M-1}\sum_{i=1}^{N}\sum_{r\neq i}\omega_{pl}{}^{-1}\omega_{tj}{}^{-1}\psi_{2rp}\left(\psi_{2r1}\psi_{2rl}-\omega_{1l}\right)\psi_{2it}\left(\psi_{2i,M-1}\psi_{2ij}-\omega_{M-1,j}\right) \\ \vdots & \ddots & \vdots \\ \sum_{p=1}^{M-1}\sum_{l=1}^{M-1}\sum_{t=1}^{M-1}\sum_{j=1}^{M-1}\sum_{i=1}^{N}\sum_{r\neq i}\omega_{pl}{}^{-1}\omega_{tj}{}^{-1}\psi_{2rp}\left(\psi_{2r,M-1}\psi_{2rl}-\omega_{M-1,l}\right)\psi_{2it}\left(\psi_{2i1}\psi_{2ij}-\omega_{1j}\right) & \cdots & \sum_{p=1}^{M-1}\sum_{l=1}^{M-1}\sum_{t=1}^{M-1}\sum_{j=1}^{M-1}\sum_{i=1}^{N}\sum_{r\neq i}\omega_{pl}{}^{-1}\omega_{tj}{}^{-1}\psi_{2rp}\left(\psi_{2r,M-1}\psi_{2rl}-\omega_{M-1,l}\right)\psi_{2it}\left(\psi_{2i,M-1}\psi_{2ij}-\omega_{M-1,j}\right) \end{array}\right)\Omega^{-1}\sigma_{V\psi_2}$$

$$+\frac{1}{N^4}\sigma_{V\psi_2}{}'\Omega^{-1}E\left(\begin{array}{ccc} \sum_{p=1}^{M-1}\sum_{l=1}^{M-1}\sum_{t=1}^{M-1}\sum_{j=1}^{M-1}\sum_{i=1}^{N}\sum_{s\neq i}\omega_{pl}{}^{-1}\omega_{tj}{}^{-1}\psi_{2st}\left(\psi_{2s1}\psi_{2sl}-\omega_{1l}\right)\psi_{2ip}\left(\psi_{2i1}\psi_{2ij}-\omega_{1j}\right) & \cdots & \sum_{p=1}^{M-1}\sum_{l=1}^{M-1}\sum_{t=1}^{M-1}\sum_{j=1}^{M-1}\sum_{i=1}^{N}\sum_{s\neq i}\omega_{pl}{}^{-1}\omega_{tj}{}^{-1}\psi_{2st}\left(\psi_{2s1}\psi_{2sl}-\omega_{1l}\right)\psi_{2ip}\left(\psi_{2i,M-1}\psi_{2ij}-\omega_{M-1,j}\right) \\ \vdots & \ddots & \vdots \\ \sum_{p=1}^{M-1}\sum_{l=1}^{M-1}\sum_{t=1}^{M-1}\sum_{j=1}^{M-1}\sum_{i=1}^{N}\sum_{s\neq i}\omega_{pl}{}^{-1}\omega_{tj}{}^{-1}\psi_{2st}\left(\psi_{2s,M-1}\psi_{2sl}-\omega_{M-1,l}\right)\psi_{2ip}\left(\psi_{2i1}\psi_{2ij}-\omega_{1j}\right) & \cdots & \sum_{p=1}^{M-1}\sum_{l=1}^{M-1}\sum_{t=1}^{M-1}\sum_{j=1}^{M-1}\sum_{i=1}^{N}\sum_{s\neq i}\omega_{pl}{}^{-1}\omega_{tj}{}^{-1}\psi_{2st}\left(\psi_{2s,M-1}\psi_{2sl}-\omega_{M-1,l}\right)\psi_{2ip}\left(\psi_{2i,M-1}\psi_{2ij}-\omega_{M-1,j}\right) \end{array}\right)\Omega^{-1}\sigma_{V\psi_2}$$

$$+\frac{1}{N^4}\sigma_{V\psi_2}{}'\Omega^{-1}E\left(\begin{array}{c} \sum_{p=1}^{M-1}\sum_{l=1}^{M-1}\sum_{t=1}^{M-1}\sum_{j=1}^{M-1}\sum_{i=1}^{N}\sum_{s\neq i}\omega_{pl}{}^{-1}\omega_{tj}{}^{-1}\psi_{2sp}\psi_{2st}\left(\psi_{2i1}\psi_{2il}-\omega_{1l}\right)\left(\psi_{2i1}\psi_{2ij}-\omega_{1j}\right) \\ \vdots \\ \sum_{p=1}^{M-1}\sum_{l=1}^{M-1}\sum_{t=1}^{M-1}\sum_{j=1}^{M-1}\sum_{i=1}^{N}\sum_{s\neq i}\omega_{pl}{}^{-1}\omega_{tj}{}^{-1}\psi_{2sp}\psi_{2st}\left(\psi_{2i,M-1}\psi_{2il}-\omega_{M-1,l}\right)\left(\psi_{2i1}\psi_{2ij}-\omega_{1j}\right) \end{array}\right.$$

$$\left.\begin{array}{lc} \cdots & \sum_{p=1}^{M-1}\sum_{l=1}^{M-1}\sum_{t=1}^{M-1}\sum_{j=1}^{M-1}\sum_{t=1}^{N}\sum_{s\neq i}\omega_{pl}{}^{-1}\omega_{tj}{}^{-1}\psi_{2sp}\psi_{2st}\left(\psi_{2i1}\psi_{2il}-\omega_{1l}\right)\left(\psi_{2i,M-1}\psi_{2ij}-\omega_{M-1,j}\right) \\ \ddots & \vdots \\ \cdots & \sum_{p=1}^{M-1}\sum_{l=1}^{M-1}\sum_{t=1}^{M-1}\sum_{j=1}^{M-1}\sum_{t=1}^{N}\sum_{s\neq i}\omega_{pl}{}^{-1}\omega_{tj}{}^{-1}\psi_{2sp}\psi_{2st}\left(\psi_{2i,M-1}\psi_{2il}-\omega_{M-1,l}\right)\left(\psi_{2i,M-1}\psi_{2ij}-\omega_{M-1,j}\right) \end{array}\right)\Omega^{-1}\sigma_{V\psi_2}$$

$$+o_p\left(N^{-2}\right)$$

$$=\frac{1}{N^2}\sigma_{V\psi_2}{}'\Omega^{-1}\left(\begin{array}{c} \sum_{p=1}^{M-1}\sum_{l=1}^{M-1}\sum_{t=1}^{M-1}\sum_{j=1}^{M-1}\omega_{pl}{}^{-1}\omega_{tj}{}^{-1}Cov\left(\psi_{2rp},\psi_{2r1}\psi_{2rl}\right)Cov\left(\psi_{2it},\psi_{2i1}\psi_{2ij}\right) \\ \vdots \\ \sum_{p=1}^{M-1}\sum_{l=1}^{M-1}\sum_{t=1}^{M-1}\sum_{j=1}^{M-1}\omega_{pl}{}^{-1}\omega_{tj}{}^{-1}Cov\left(\psi_{2rp},\psi_{2r,M-1}\psi_{2rl}\right)Cov\left(\psi_{2it},\psi_{2i1}\psi_{2ij}\right) \end{array}\right.$$

$$\left.\begin{array}{lc} \cdots & \sum_{p=1}^{M-1}\sum_{l=1}^{M-1}\sum_{t=1}^{M-1}\sum_{j=1}^{M-1}\omega_{pl}{}^{-1}\omega_{tj}{}^{-1}Cov\left(\psi_{2rp},\psi_{2r1}\psi_{2rl}\right)Cov\left(\psi_{2it},\psi_{2i,M-1}\psi_{2ij}\right) \\ \ddots & \vdots \\ \cdots & \sum_{p=1}^{M-1}\sum_{l=1}^{M-1}\sum_{t=1}^{M-1}\sum_{j=1}^{M-1}\omega_{pl}{}^{-1}\omega_{tj}{}^{-1}Cov\left(\psi_{2rp},\psi_{2r,M-1}\psi_{2rl}\right)Cov\left(\psi_{2it},\psi_{2i,M-1}\psi_{2ij}\right) \end{array}\right)\Omega^{-1}\sigma_{V\psi_2}$$

$$+\frac{1}{N^2}\sigma_{V\psi_2}{}'\Omega^{-1}\begin{pmatrix} \sum_{p=1}^{M-1}\sum_{l=1}^{M-1}\sum_{t=1}^{M-1}\sum_{j=1}^{M-1}\omega_{pl}{}^{-1}\omega_{tj}{}^{-1}Cov\left(\psi_{2st},\psi_{2s1}\psi_{2sl}\right)Cov\left(\psi_{2ip},\psi_{2i1}\psi_{2ij}\right) & \cdots & \sum_{p=1}^{M-1}\sum_{l=1}^{M-1}\sum_{t=1}^{M-1}\sum_{j=1}^{M-1}\omega_{pl}{}^{-1}\omega_{tj}{}^{-1}Cov\left(\psi_{2st},\psi_{2s1}\psi_{2sl}\right)Cov\left(\psi_{2ip},\psi_{2i,M-1}\psi_{2ij}\right) \\ \vdots & \ddots & \vdots \\ \sum_{p=1}^{M-1}\sum_{l=1}^{M-1}\sum_{t=1}^{M-1}\sum_{j=1}^{M-1}\omega_{pl}{}^{-1}\omega_{tj}{}^{-1}Cov\left(\psi_{2st},\psi_{2s,M-1}\psi_{2sl}\right)Cov\left(\psi_{2ip},\psi_{2i1}\psi_{2ij}\right) & \cdots & \sum_{p=1}^{M-1}\sum_{l=1}^{M-1}\sum_{t=1}^{M-1}\sum_{j=1}^{M-1}\omega_{pl}{}^{-1}\omega_{tj}{}^{-1}Cov\left(\psi_{2st},\psi_{2s,M-1}\psi_{2sl}\right)Cov\left(\psi_{2ip},\psi_{2i,M-1}\psi_{2ij}\right) \end{pmatrix}\Omega^{-1}\sigma_{V\psi_2}$$

$$+\frac{1}{N^2}\sigma_{V\psi_2}{}'\Omega^{-1}\begin{pmatrix} \sum_{p=1}^{M-1}\sum_{l=1}^{M-1}\sum_{t=1}^{M-1}\sum_{j=1}^{M-1}\omega_{pl}{}^{-1}\omega_{tj}{}^{-1}Cov\left(\psi_{2sp},\psi_{2st}\right)Cov\left(\psi_{2i1}\psi_{2il},\psi_{2i1}\psi_{2ij}\right) & \cdots & \sum_{p=1}^{M-1}\sum_{l=1}^{M-1}\sum_{t=1}^{M-1}\sum_{j=1}^{M-1}\omega_{pl}{}^{-1}\omega_{tj}{}^{-1}Cov\left(\psi_{2sp},\psi_{2st}\right)Cov\left(\psi_{2i1}\psi_{2il},\psi_{2i,M-1}\psi_{2ij}\right) \\ \vdots & \ddots & \vdots \\ \sum_{p=1}^{M-1}\sum_{l=1}^{M-1}\sum_{t=1}^{M-1}\sum_{j=1}^{M-1}\omega_{pl}{}^{-1}\omega_{tj}{}^{-1}Cov\left(\psi_{2sp},\psi_{2st}\right)Cov\left(\psi_{2i,M-1}\psi_{2il},\psi_{2i1}\psi_{2ij}\right) & \cdots & \sum_{p=1}^{M-1}\sum_{l=1}^{M-1}\sum_{t=1}^{M-1}\sum_{j=1}^{M-1}\omega_{pl}{}^{-1}\omega_{tj}{}^{-1}Cov\left(\psi_{2sp},\psi_{2st}\right)Cov\left(\psi_{2i,M-1}\psi_{2il},\psi_{2i,M-1}\psi_{2ij}\right) \end{pmatrix}\Omega^{-1}\sigma_{V\psi_2}$$

$$+o_p\left(N^{-2}\right)$$

$$E\left(R_{n2}{}^2\right)$$

$$=E\Bigg(\frac{1}{N^2}\Bigg(\sum_{p=1}^{M-1}\sum_{l=1}^{M-1}\sum_{k=1}^{N}\sum_{r=1}^{N}\omega_{pl}{}^{-1}\psi_{2kp}\left(\left(V_r-\tau\right)\psi_{2rl}-\sigma_{V\psi_2 l}\right)\Bigg)\cdot\frac{1}{N^2}\Bigg(\sum_{t=1}^{M-1}\sum_{j=1}^{M-1}\sum_{s=1}^{N}\sum_{i=1}^{N}\omega_{tj}{}^{-1}\psi_{2st}\left(\left(V_i-\tau\right)\psi_{2ij}-\sigma_{V\psi_2 j}\right)$$

$$=\frac{1}{N^4}E\left(\sum_{p=1}^{M-1}\sum_{l=1}^{M-1}\sum_{t=1}^{M-1}\sum_{j=1}^{M-1}\sum_{k=1}^{N}\sum_{r=1}^{N}\sum_{s=1}^{N}\sum_{i=1}^{N}\omega_{pl}{}^{-1}\omega_{tj}{}^{-1}\psi_{2kp}\psi_{2st}\left(\left(V_r-\tau\right)\psi_{2rl}-\sigma_{V\psi_2 l}\right)\left(\left(V_i-\tau\right)\psi_{2ij}-\sigma_{V\psi_2 j}\right)\right)$$

For the same reason as mentioned above, we consider the three cases when $(k=r, s=i, r\neq i)$, $(k=i, r=s, s\neq i)$, and $(k=s, r=i, s\neq i)$. The equation above becomes

$$=\frac{1}{N^4}E\left(\sum_{p=1}^{M-1}\sum_{l=1}^{M-1}\sum_{t=1}^{M-1}\sum_{j=1}^{M-1}\sum_{i=1}^{N}\sum_{r\neq i}\omega_{pl}{}^{-1}\omega_{tj}{}^{-1}\psi_{2rp}\left(\left(V_r-\tau\right)\psi_{2rl}-\sigma_{V\psi_2 l}\right)\psi_{2it}\left(\left(V_i-\tau\right)\psi_{2ij}-\sigma_{V\psi_2 J}\right)\right)$$

$$+\frac{1}{N^4}E\left(\sum_{p=1}^{M-1}\sum_{l=1}^{M-1}\sum_{t=1}^{M-1}\sum_{j=1}^{M-1}\sum_{i=1}^{N}\sum_{s\neq i}\omega_{pl}{}^{-1}\omega_{tj}{}^{-1}\psi_{2st}\left(\left(V_s-\tau\right)\psi_{2sl}-\sigma_{V\psi_2 l}\right)\psi_{2ip}\left(\left(V_i-\tau\right)\psi_{2ij}-\sigma_{V\psi_2 j}\right)\right)$$

$$+\frac{1}{N^4}E\left(\sum_{p=1}^{M-1}\sum_{l=1}^{M-1}\sum_{t=1}^{M-1}\sum_{j=1}^{M-1}\sum_{i=1}^{N}\sum_{s\neq i}\omega_{pl}{}^{-1}\omega_{tj}{}^{-1}\psi_{2sp}\psi_{2st}\left(\left(V_i-\tau\right)\psi_{2il}-\sigma_{V\psi_2 l}\right)\left(\left(V_i-\tau\right)\psi_{2ij}-\sigma_{V\psi_2 J}\right)\right)$$

$$+o_p\left(N^{-2}\right)$$

$$=\frac{1}{N^2}\left(\sum_{p=1}^{M-1}\sum_{l=1}^{M-1}\sum_{t=1}^{M-1}\sum_{j=1}^{M-1}\omega_{pl}{}^{-1}\omega_{tj}{}^{-1}Cov\left(\psi_{2ip},\left(V_i-\tau\right)\psi_{2il}\right)Cov\left(\psi_{2it},\left(V_i-\tau\right)\psi_{2ij}\right)\right)$$

$$+\frac{1}{N^2}\left(\sum_{p=1}^{M-1}\sum_{l=1}^{M-1}\sum_{t=1}^{M-1}\sum_{j=1}^{M-1}\omega_{pl}{}^{-1}\omega_{tj}{}^{-1}Cov\left(\psi_{2it},\left(V_i-\tau\right)\psi_{2il}\right)Cov\left(\psi_{2ip},\left(V_i-\tau\right)\psi_{2ij}\right)\right)$$

$$+\frac{1}{N^2}\left(\sum_{p=1}^{M-1}\sum_{l=1}^{M-1}\sum_{t=1}^{M-1}\sum_{j=1}^{M-1}\omega_{pl}{}^{-1}\omega_{tj}{}^{-1}Cov\left(\psi_{2ip},\psi_{2it}\right)Cov\left(\left(V_i-\tau\right)\psi_{2il},\left(V_i-\tau\right)_i\psi_{2ij}\right)\right)$$

$$+o_p\left(N^{-2}\right)$$

$$E\left(T_{N1}S_{N1}\right)$$

$$
=-\frac{1}{N}E\left(\left(\sum_{k=1}^{N}(V_k-\tau)\right)\cdot\frac{1}{N^3}\begin{pmatrix}\sum_{p=1}^{M-1}\sum_{l=1}^{M-1}\sum_{t=1}^{M-1}\sum_{j=1}^{M-1}\sum_{r=1}^{N}\sum_{s=1}^{N}\sum_{i=1}^{N}\psi_{2rp}\omega_{pl}{}^{-1}\left(\psi_{2st}\psi_{2sl}-\omega_{tl}\right)\omega_{tj}{}^{-1}\left(\psi_{2ij}\psi_{2i1}-\omega_{1j}\right)\\ \vdots \\ \sum_{p=1}^{M-1}\sum_{l=1}^{M-1}\sum_{t=1}^{M-1}\sum_{j=1}^{M-1}\sum_{r=1}^{N}\sum_{s=1}^{N}\sum_{i=1}^{N}\psi_{2rp}\omega_{pl}{}^{-1}\left(\psi_{2st}\psi_{2sl}-\omega_{tl}\right)\omega_{tj}{}^{-1}\left(\psi_{2ij}\psi_{2i,M-1}-\omega_{M-1,j}\right)\end{pmatrix}'\right)
$$

$$
\times\Omega^{-1}\sigma_{V\psi_2}
$$

$$
=-\frac{1}{N^4}E\begin{pmatrix}\sum_{p=1}^{M-1}\sum_{l=1}^{M-1}\sum_{t=1}^{M-1}\sum_{j=1}^{M-1}\sum_{k=1}^{N}\sum_{r=1}^{N}\sum_{s=1}^{N}\sum_{i=1}^{N}\omega_{pl}{}^{-1}\omega_{tj}{}^{-1}(V_k-\tau)\psi_{2rp}\left(\psi_{2st}\psi_{2sl}-\omega_{tl}\right)\left(\psi_{2ij}\psi_{2i1}-\omega_{1j}\right)\\ \vdots \\ \sum_{p=1}^{M-1}\sum_{l=1}^{M-1}\sum_{t=1}^{M-1}\sum_{j=1}^{M-1}\sum_{k=1}^{N}\sum_{r=1}^{N}\sum_{s=1}^{N}\sum_{i=1}^{N}\omega_{pl}{}^{-1}\omega_{tj}{}^{-1}(V_k-\tau)\psi_{2rp}\left(\psi_{2st}\psi_{2sl}-\omega_{tl}\right)\left(\psi_{2ij}\psi_{2i,M-1}-\omega_{M-1,j}\right)\end{pmatrix}'\Omega^{-1}\sigma_{V\psi}
$$

For the same reason as mentioned above, we consider the three cases when $(k=r,s=i,r\neq i)$, $(k=i,r=s,s\neq i)$, and $(k=s,r=i,s\neq i)$. The equation above becomes

$$
=-\frac{1}{N^4}E\begin{pmatrix}\sum_{p=1}^{M-1}\sum_{l=1}^{M-1}\sum_{t=1}^{M-1}\sum_{j=1}^{M-1}\sum_{i=1}^{N}\sum_{r\neq i}\omega_{pl}{}^{-1}\omega_{tj}{}^{-1}(V_r-\tau)\psi_{2rp}\left(\psi_{2it}\psi_{2il}-\omega_{tl}\right)\left(\psi_{2ij}\psi_{2i1}-\omega_{1j}\right)\\ \vdots \\ \sum_{p=1}^{M-1}\sum_{l=1}^{M-1}\sum_{t=1}^{M-1}\sum_{j=1}^{M-1}\sum_{i=1}^{N}\sum_{r\neq i}\omega_{pl}{}^{-1}\omega_{tj}{}^{-1}(V_r-\tau)\psi_{2rp}\left(\psi_{2it}\psi_{2il}-\omega_{tl}\right)\left(\psi_{2ij}\psi_{2i,M-1}-\omega_{M-1,j}\right)\end{pmatrix}'\Omega^{-1}\sigma_{V\psi_2}
$$

$$
-\frac{1}{N^4}E\begin{pmatrix}\sum_{p=1}^{M-1}\sum_{l=1}^{M-1}\sum_{t=1}^{M-1}\sum_{j=1}^{M-1}\sum_{i=1}^{N}\sum_{s\neq i}\omega_{pl}{}^{-1}\omega_{tj}{}^{-1}\psi_{2sp}\left(\psi_{2st}\psi_{2sl}-\omega_{tl}\right)(V_i-\tau)\left(\psi_{2ij}\psi_{2i1}-\omega_{1j}\right)\\ \vdots \\ \sum_{p=1}^{M-1}\sum_{l=1}^{M-1}\sum_{t=1}^{M-1}\sum_{j=1}^{M-1}\sum_{i=1}^{N}\sum_{s\neq i}\omega_{pl}{}^{-1}\omega_{tj}{}^{-1}\psi_{2sp}\left(\psi_{2st}\psi_{2sl}-\omega_{tl}\right)(V_i-\tau)\left(\psi_{2ij}\psi_{2i,M-1}-\omega_{M-1,j}\right)\end{pmatrix}'\Omega^{-1}\sigma_{V\psi_2}
$$

$$-\frac{1}{N^4}E\begin{pmatrix} \sum_{p=1}^{M-1}\sum_{l=1}^{M-1}\sum_{t=1}^{M-1}\sum_{j=1}^{M-1}\sum_{i=1}^{N}\sum_{s\neq i}\omega_{pl}^{-1}\omega_{tj}^{-1}\left(V_s-\tau\right)\left(\psi_{2st}\psi_{2sl}-\omega_{tl}\right)\psi_{2ip}\left(\psi_{2ij}\psi_{2i1}-\omega_{1j}\right) \\ \vdots \\ \sum_{p=1}^{M-1}\sum_{l=1}^{M-1}\sum_{t=1}^{M-1}\sum_{j=1}^{M-1}\sum_{i=1}^{N}\sum_{s\neq i}\omega_{pl}^{-1}\omega_{tj}^{-1}\left(V_s-\tau\right)\left(\psi_{2st}\psi_{2sl}-\omega_{tl}\right)\psi_{2ip}\left(\psi_{2ij}\psi_{2i,M-1}-\omega_{M-1,j}\right) \end{pmatrix}'\Omega^{-1}\sigma_{V\psi_2}$$

$$+o_p\left(N^{-2}\right)$$

$$=-\frac{1}{N^2}\begin{pmatrix} \sum_{p=1}^{M-1}\sum_{l=1}^{M-1}\sum_{t=1}^{M-1}\sum_{j=1}^{M-1}\omega_{pl}^{-1}\omega_{tj}^{-1}Cov\left(\left(V_i-\tau\right),\psi_{2ip}\right)Cov\left(\psi_{2it}\psi_{2il},\psi_{2ij}\psi_{2i1}\right) \\ \vdots \\ \sum_{p=1}^{M-1}\sum_{l=1}^{M-1}\sum_{t=1}^{M-1}\sum_{j=1}^{M-1}\omega_{pl}^{-1}\omega_{tj}^{-1}Cov\left(\left(V_i-\tau\right),\psi_{2ip}\right)Cov\left(\psi_{2it}\psi_{2il},\psi_{2ij}\psi_{2i,M-1}\right) \end{pmatrix}'\Omega^{-1}\sigma_{V\psi_2}$$

$$-\frac{1}{N^2}\begin{pmatrix} \sum_{p=1}^{M-1}\sum_{l=1}^{M-1}\sum_{t=1}^{M-1}\sum_{j=1}^{M-1}\omega_{pl}^{-1}\omega_{tj}^{-1}Cov\left(\psi_{2ip},\psi_{2it}\psi_{2il}\right)Cov\left(\left(V_i-\tau\right),\psi_{2ij}\psi_{2i1}\right) \\ \vdots \\ \sum_{p=1}^{M-1}\sum_{l=1}^{M-1}\sum_{t=1}^{M-1}\sum_{j=1}^{M-1}\omega_{pl}^{-1}\omega_{tj}^{-1}Cov\left(\psi_{2ip},\psi_{2it}\psi_{2il}\right)Cov\left(\left(V_i-\tau\right),\psi_{2ij}\psi_{2i,M-1}\right) \end{pmatrix}'\Omega^{-1}\sigma_{V\psi_2}$$

$$-\frac{1}{N^2}\begin{pmatrix} \sum_{p=1}^{M-1}\sum_{l=1}^{M-1}\sum_{t=1}^{M-1}\sum_{j=1}^{M-1}\omega_{pl}^{-1}\omega_{tj}^{-1}Cov\left(\left(V_i-\tau\right),\psi_{2it}\psi_{2il}\right)Cov\left(\psi_{2ip},\psi_{2ij}\psi_{2i1}\right) \\ \vdots \\ \sum_{p=1}^{M-1}\sum_{l=1}^{M-1}\sum_{t=1}^{M-1}\sum_{j=1}^{M-1}\omega_{pl}^{-1}\omega_{tj}^{-1}Cov\left(\left(V_i-\tau\right),\psi_{2it}\psi_{2il}\right)Cov\left(\psi_{2ip},\psi_{2ij}\psi_{2i,M-1}\right) \end{pmatrix}'\Omega^{-1}\sigma_{V\psi_2}$$

$$+o_p\left(N^{-2}\right)$$

$$E\left(T_{N2}S_{N1}\right)$$

$$=\frac{1}{N}\sigma_{V\psi_2}{}'\Omega^{-1}E\left(\left(\sum_{k=1}^{N}\begin{pmatrix}\psi_{2k1}\\ \vdots \\ \psi_{2k,M-1}\end{pmatrix}\right)\frac{1}{N^3}\begin{pmatrix}\sum_{p=1}^{M-1}\sum_{l=1}^{M-1}\sum_{t=1}^{M-1}\sum_{j=1}^{M-1}\sum_{r=1}^{N}\sum_{s=1}^{N}\sum_{i=1}^{N}\psi_{2rp}\omega_{pl}{}^{-1}\left(\psi_{2st}\psi_{2sl}-\omega_{tl}\right)\omega_{tj}{}^{-1}\left(\psi_{2ij}\psi_{2i1}-\omega_{1j}\right)\\ \vdots \\ \sum_{p=1}^{M-1}\sum_{l=1}^{M-1}\sum_{t=1}^{M-1}\sum_{j=1}^{M-1}\sum_{r=1}^{N}\sum_{s=1}^{N}\sum_{i=1}^{N}\psi_{2rp}\omega_{pl}{}^{-1}\left(\psi_{2st}\psi_{2sl}-\omega_{tl}\right)\omega_{tj}{}^{-1}\left(\psi_{2ij}\psi_{2i,M-1}-\omega_{M-1,j}\right)\end{pmatrix}'\right)$$

$$\times\Omega^{-1}\sigma_{V\psi_2}$$

$$=\frac{1}{N^4}\sigma_{V\psi_2}{}'\Omega^{-1}E\begin{pmatrix}\sum_{p=1}^{M-1}\sum_{l=1}^{M-1}\sum_{t=1}^{M-1}\sum_{j=1}^{M-1}\sum_{k=1}^{N}\sum_{r=1}^{N}\sum_{s=1}^{N}\sum_{i=1}^{N}\omega_{pl}{}^{-1}\omega_{tj}{}^{-1}\psi_{2k1}\psi_{2rp}\left(\psi_{2st}\psi_{2sl}-\omega_{tl}\right)\left(\psi_{2ij}\psi_{2i1}-\omega_{1j}\right) & \cdots & \sum_{p=1}^{M-1}\sum_{l=1}^{M-1}\sum_{t=1}^{M-1}\sum_{j=1}^{M-1}\sum_{k=1}^{N}\sum_{r=1}^{N}\sum_{s=1}^{N}\sum_{i=1}^{N}\omega_{pl}{}^{-1}\omega_{tj}{}^{-1}\psi_{2k1}\psi_{2rp}\left(\psi_{2st}\psi_{2sl}-\omega_{tl}\right)\left(\psi_{2ij}\psi_{2i,M-1}-\omega_{M-1,j}\right)\\ \vdots & \ddots & \vdots \\ \sum_{p=1}^{M-1}\sum_{l=1}^{M-1}\sum_{t=1}^{M-1}\sum_{j=1}^{M-1}\sum_{k=1}^{N}\sum_{r=1}^{N}\sum_{s=1}^{N}\sum_{i=1}^{N}\omega_{pl}{}^{-1}\omega_{tj}{}^{-1}\psi_{2k,M-1}\psi_{2rp}\left(\psi_{2st}\psi_{2sl}-\omega_{tl}\right)\left(\psi_{2ij}\psi_{2i1}-\omega_{1j}\right) & \cdots & \sum_{p=1}^{M-1}\sum_{l=1}^{M-1}\sum_{t=1}^{M-1}\sum_{j=1}^{M-1}\sum_{k=1}^{N}\sum_{r=1}^{N}\sum_{s=1}^{N}\sum_{i=1}^{N}\omega_{pl}{}^{-1}\omega_{tj}{}^{-1}\psi_{2k,M-1}\psi_{2rp}\left(\psi_{2st}\psi_{2sl}-\omega_{tl}\right)\left(\psi_{2ij}\psi_{2i,M-1}-\omega_{M-1,j}\right)\end{pmatrix}\Omega^{-1}\sigma_{V\psi_2}$$

For the same reason as mentioned above, we consider the three cases when $\left(k=r,s=i,r\neq i\right)$, $\left(k=i,r=s,s\neq i\right)$, and $\left(k=s,r=i,s\neq i\right)$. The equation above becomes

$$=\frac{1}{N^4}\sigma_{V\psi_2}{}'\Omega^{-1}E\left(\begin{matrix}\sum_{p=1}^{M-1}\sum_{l=1}^{M-1}\sum_{t=1}^{M-1}\sum_{j=1}^{M-1}\sum_{i=1}^{N}\sum_{r\neq i}\omega_{pl}{}^{-1}\omega_{tj}{}^{-1}\psi_{2r1}\psi_{2rp}\left(\psi_{2it}\psi_{2il}-\omega_{tl}\right)\left(\psi_{2ij}\psi_{2i1}-\omega_{1j}\right)\\ \vdots\\ \sum_{p=1}^{M-1}\sum_{l=1}^{M-1}\sum_{t=1}^{M-1}\sum_{j=1}^{M-1}\sum_{i=1}^{N}\sum_{r\neq i}\omega_{pl}{}^{-1}\omega_{tj}{}^{-1}\psi_{2r,M-1}\psi_{2rp}\left(\psi_{2it}\psi_{2il}-\omega_{tl}\right)\left(\psi_{2ij}\psi_{2i1}-\omega_{1j}\right)\end{matrix}\right.$$

$$\left.\begin{matrix}\cdots & \sum_{p=1}^{M-1}\sum_{l=1}^{M-1}\sum_{t=1}^{M-1}\sum_{j=1}^{M-1}\sum_{i=1}^{N}\sum_{r\neq i}\omega_{pl}{}^{-1}\omega_{tj}{}^{-1}\psi_{2r1}\psi_{2rp}\left(\psi_{2it}\psi_{2il}-\omega_{tl}\right)\left(\psi_{2ij}\psi_{2i,M-1}-\omega_{M-1,j}\right)\\ \ddots & \vdots\\ \cdots & \sum_{p=1}^{M-1}\sum_{l=1}^{M-1}\sum_{t=1}^{M-1}\sum_{j=1}^{M-1}\sum_{i=1}^{N}\sum_{r\neq i}\omega_{pl}{}^{-1}\omega_{tj}{}^{-1}\psi_{2r,M-1}\psi_{2rp}\left(\psi_{2it}\psi_{2il}-\omega_{tl}\right)\left(\psi_{2ij}\psi_{2i,M-1}-\omega_{M-1,j}\right)\end{matrix}\right)\Omega^{-1}\sigma_{V\psi_2}$$

$$+\frac{1}{N^4}\sigma_{V\psi_2}{}'\Omega^{-1}E\left(\begin{matrix}\sum_{p=1}^{M-1}\sum_{l=1}^{M-1}\sum_{t=1}^{M-1}\sum_{j=1}^{M-1}\sum_{i=1}^{N}\sum_{s\neq i}\omega_{pl}{}^{-1}\omega_{tj}{}^{-1}\psi_{2sp}\left(\psi_{2st}\psi_{2sl}-\omega_{tl}\right)\psi_{2i1}\left(\psi_{2ij}\psi_{2i1}-\omega_{1j}\right)\\ \vdots\\ \sum_{p=1}^{M-1}\sum_{l=1}^{M-1}\sum_{t=1}^{M-1}\sum_{j=1}^{M-1}\sum_{i=1}^{N}\sum_{s\neq i}\omega_{pl}{}^{-1}\omega_{tj}{}^{-1}\psi_{2sp}\left(\psi_{2st}\psi_{2sl}-\omega_{tl}\right)\psi_{2i,M-1}\left(\psi_{2ij}\psi_{2i1}-\omega_{1j}\right)\end{matrix}\right.$$

$$\left.\begin{matrix}\cdots & \sum_{p=1}^{M-1}\sum_{l=1}^{M-1}\sum_{t=1}^{M-1}\sum_{j=1}^{M-1}\sum_{i=1}^{N}\sum_{s\neq i}\omega_{pl}{}^{-1}\omega_{tj}{}^{-1}\psi_{2sp}\left(\psi_{2st}\psi_{2sl}-\omega_{tl}\right)\psi_{2i1}\left(\psi_{2ij}\psi_{2i,M-1}-\omega_{M-1,j}\right)\\ \ddots & \vdots\\ \cdots & \sum_{p=1}^{M-1}\sum_{l=1}^{M-1}\sum_{t=1}^{M-1}\sum_{j=1}^{M-1}\sum_{i=1}^{N}\sum_{s\neq i}\omega_{pl}{}^{-1}\omega_{tj}{}^{-1}\psi_{2sp}\left(\psi_{2st}\psi_{2sl}-\omega_{tl}\right)\psi_{2i,M-1}\left(\psi_{2ij}\psi_{2i,M-1}-\omega_{M-1,j}\right)\end{matrix}\right)\Omega^{-1}\sigma_{V\psi_2}$$

$$+\frac{1}{N^4}\sigma_{V\psi_2}{}'\Omega^{-1}E\left(\begin{matrix}\sum_{p=1}^{M-1}\sum_{l=1}^{M-1}\sum_{t=1}^{M-1}\sum_{j=1}^{M-1}\sum_{i=1}^{N}\sum_{s\neq i}\omega_{pl}{}^{-1}\omega_{tj}{}^{-1}\psi_{2s1}\left(\psi_{2st}\psi_{2sl}-\omega_{tl}\right)\psi_{2ip}\left(\psi_{2ij}\psi_{2i1}-\omega_{1j}\right)\\ \vdots\\ \sum_{p=1}^{M-1}\sum_{l=1}^{M-1}\sum_{t=1}^{M-1}\sum_{j=1}^{M-1}\sum_{i=1}^{N}\sum_{s\neq i}\omega_{pl}{}^{-1}\omega_{tj}{}^{-1}\psi_{2s,M-1}\left(\psi_{2st}\psi_{2sl}-\omega_{tl}\right)\psi_{2ip}\left(\psi_{2ij}\psi_{2i1}-\omega_{1j}\right)\end{matrix}\right.$$

$$\left.\begin{matrix}\cdots & \sum_{p=1}^{M-1}\sum_{l=1}^{M-1}\sum_{t=1}^{M-1}\sum_{j=1}^{M-1}\sum_{i=1}^{N}\sum_{s\neq i}\omega_{pl}{}^{-1}\omega_{tj}{}^{-1}\psi_{2s1}\left(\psi_{2st}\psi_{2sl}-\omega_{tl}\right)\psi_{2ip}\left(\psi_{2ij}\psi_{2i,M-1}-\omega_{M-1,j}\right)\\ \ddots & \vdots\\ \cdots & \sum_{p=1}^{M-1}\sum_{l=1}^{M-1}\sum_{t=1}^{M-1}\sum_{j=1}^{M-1}\sum_{i=1}^{N}\sum_{s\neq i}\omega_{pl}{}^{-1}\omega_{tj}{}^{-1}\psi_{2s,M-1}\left(\psi_{2st}\psi_{2sl}-\omega_{tl}\right)\psi_{2ip}\left(\psi_{2ij}\psi_{2i,M-1}-\omega_{M-1,j}\right)\end{matrix}\right)\Omega^{-1}\sigma_{V\psi_2}$$

$$+o_p\left(N^{-2}\right)$$

$$
=\frac{1}{N^2}\sigma_{V\psi_2}{}'\Omega^{-1}\begin{pmatrix}
\sum_{p=1}^{M-1}\sum_{l=1}^{M-1}\sum_{t=1}^{M-1}\sum_{j=1}^{M-1}\omega_{pl}{}^{-1}\omega_{tj}{}^{-1}Cov\left(\psi_{2i1},\psi_{2ip}\right)Cov\left(\psi_{2it}\psi_{2il},\psi_{2ij}\psi_{2i1}\right) & \cdots & \sum_{p=1}^{M-1}\sum_{l=1}^{M-1}\sum_{t=1}^{M-1}\sum_{j=1}^{M-1}\omega_{pl}{}^{-1}\omega_{tj}{}^{-1}Cov\left(\psi_{2i1},\psi_{2ip}\right)Cov\left(\psi_{2it}\psi_{2il},\psi_{2ij}\psi_{2i,M-1}\right) \\
\vdots & \ddots & \vdots \\
\sum_{p=1}^{M-1}\sum_{l=1}^{M-1}\sum_{t=1}^{M-1}\sum_{j=1}^{M-1}\omega_{pl}{}^{-1}\omega_{tj}{}^{-1}Cov\left(\psi_{2i,M-1},\psi_{2ip}\right)Cov\left(\psi_{2it}\psi_{2il},\psi_{2ij}\psi_{2i1}\right) & \cdots & \sum_{p=1}^{M-1}\sum_{l=1}^{M-1}\sum_{t=1}^{M-1}\sum_{j=1}^{M-1}\omega_{pl}{}^{-1}\omega_{tj}{}^{-1}Cov\left(\psi_{2i,M-1},\psi_{2ip}\right)Cov\left(\psi_{2it}\psi_{2il},\psi_{2ij}\psi_{2i,M-1}\right)
\end{pmatrix}\Omega^{-1}\sigma_{V\psi_2}
$$

$$
+\frac{1}{N^2}\sigma_{V\psi_2}{}'\Omega^{-1}\begin{pmatrix}
\sum_{p=1}^{M-1}\sum_{l=1}^{M-1}\sum_{t=1}^{M-1}\sum_{j=1}^{M-1}\omega_{pl}{}^{-1}\omega_{tj}{}^{-1}Cov\left(\psi_{2ip},\psi_{2it}\psi_{2il}\right)Cov\left(\psi_{2i1},\psi_{2ij}\psi_{2i1}\right) & \cdots & \sum_{p=1}^{M-1}\sum_{l=1}^{M-1}\sum_{t=1}^{M-1}\sum_{j=1}^{M-1}\omega_{pl}{}^{-1}\omega_{tj}{}^{-1}Cov\left(\psi_{2ip},\psi_{2it}\psi_{2il}\right)Cov\left(\psi_{2i1},\psi_{2ij}\psi_{2i,M-1}\right) \\
\vdots & \ddots & \vdots \\
\sum_{p=1}^{M-1}\sum_{l=1}^{M-1}\sum_{t=1}^{M-1}\sum_{j=1}^{M-1}\omega_{pl}{}^{-1}\omega_{tj}{}^{-1}Cov\left(\psi_{2ip},\psi_{2it}\psi_{2il}\right)Cov\left(\psi_{2i,M-1},\psi_{2ij}\psi_{2i1}\right) & \cdots & \sum_{p=1}^{M-1}\sum_{l=1}^{M-1}\sum_{t=1}^{M-1}\sum_{j=1}^{M-1}\omega_{pl}{}^{-1}\omega_{tj}{}^{-1}Cov\left(\psi_{2ip},\psi_{2it}\psi_{2il}\right)Cov\left(\psi_{2i,M-1},\psi_{2ij}\psi_{2i,M-1}\right)
\end{pmatrix}\Omega^{-1}\sigma_{V\psi_2}
$$

$$+\frac{1}{N^2}\sigma_{V\psi_2}'\Omega^{-1}\begin{pmatrix} \sum_{p=1}^{M-1}\sum_{l=1}^{M-1}\sum_{t=1}^{M-1}\sum_{j=1}^{M-1}\omega_{pl}^{-1}\omega_{tj}^{-1}Cov(\psi_{2i1},\psi_{2it}\psi_{2il})Cov(\psi_{2ip},\psi_{2ij}\psi_{2i1}) & \cdots & \sum_{p=1}^{M-1}\sum_{l=1}^{M-1}\sum_{t=1}^{M-1}\sum_{j=1}^{M-1}\omega_{pl}^{-1}\omega_{tj}^{-1}Cov(\psi_{2i1},\psi_{2it}\psi_{2il})Cov(\psi_{2ip},\psi_{2ij}\psi_{2i,M-1}) \\ \vdots & \ddots & \vdots \\ \sum_{p=1}^{M-1}\sum_{l=1}^{M-1}\sum_{t=1}^{M-1}\sum_{j=1}^{M-1}\omega_{pl}^{-1}\omega_{tj}^{-1}Cov(\psi_{2i,M-1},\psi_{2it}\psi_{2il})Cov(\psi_{2ip},\psi_{2ij}\psi_{2i1}) & \cdots & \sum_{p=1}^{M-1}\sum_{l=1}^{M-1}\sum_{t=1}^{M-1}\sum_{j=1}^{M-1}\omega_{pl}^{-1}\omega_{tj}^{-1}Cov(\psi_{2i,M-1},\psi_{2it}\psi_{2il})Cov(\psi_{2ip},\psi_{2ij}\psi_{2i,M-1}) \end{pmatrix}\Omega^{-1}\sigma_{V\psi_2}$$

$$+o_p\left(N^{-2}\right)$$

$$E(T_{N1}S_{N2})$$

$$=\frac{1}{N}E\left(\left(\sum_{k=1}^{N}(V_k-\tau)\right)\cdot\frac{1}{N^3}\left(\sum_{p=1}^{M-1}\sum_{l=1}^{M-1}\sum_{t=1}^{M-1}\sum_{j=1}^{M-1}\sum_{r=1}^{N}\sum_{s=1}^{N}\sum_{i=1}^{N}\psi_{2rp}\omega_{pl}^{-1}(\psi_{2st}\psi_{2sl}-\omega_{tl})\omega_{tj}^{-1}\left((V_i-\tau)\psi_{2ij}-\sigma_{V\psi_2j}\right)\right)\right)$$

$$=\frac{1}{N^4}E\left(\sum_{p=1}^{M-1}\sum_{l=1}^{M-1}\sum_{t=1}^{M-1}\sum_{j=1}^{M-1}\sum_{k=1}^{N}\sum_{r=1}^{N}\sum_{s=1}^{N}\sum_{i=1}^{N}\omega_{pl}^{-1}\omega_{tj}^{-1}(V_k-\tau)\psi_{2rp}(\psi_{2st}\psi_{2sl}-\omega_{tl})\left((V_i-\tau)\psi_{2ij}-\sigma_{V\psi_2j}\right)\right)$$

For the same reason as mentioned above, we consider the three cases when $(k=r, s=i, r\neq i)$, $(k=i, r=s, s\neq i)$, and $(k=s, r=i, s\neq i)$. The equation above becomes

$$=\frac{1}{N^4}E\left(\sum_{p=1}^{M-1}\sum_{l=1}^{M-1}\sum_{t=1}^{M-1}\sum_{j=1}^{M-1}\sum_{i=1}^{N}\sum_{r\neq i}\omega_{pl}^{-1}\omega_{tj}^{-1}(V_r-\tau)\psi_{2rp}(\psi_{2it}\psi_{2il}-\omega_{tl})\left((V_i-\tau)\psi_{2ij}-\sigma_{V\psi_2j}\right)\right)$$

$$+\frac{1}{N^4}E\left(\sum_{p=1}^{M-1}\sum_{l=1}^{M-1}\sum_{t=1}^{M-1}\sum_{j=1}^{M-1}\sum_{i=1}^{N}\sum_{s\neq i}\omega_{pl}^{-1}\omega_{tj}^{-1}\psi_{2sp}(\psi_{2st}\psi_{2sl}-\omega_{tl})(V_i-\tau)\left((V_i-\tau)\psi_{2ij}-\sigma_{V\psi_2j}\right)\right)$$

$$+\frac{1}{N^4}E\left(\sum_{p=1}^{M-1}\sum_{l=1}^{M-1}\sum_{t=1}^{M-1}\sum_{j=1}^{M-1}\sum_{i=1}^{N}\sum_{s\neq i}\omega_{pl}{}^{-1}\omega_{tj}{}^{-1}\left(V_s-\tau\right)\left(\psi_{2st}\psi_{2sl}-\omega_{tl}\right)\psi_{2sp}\left(\left(V_i-\tau\right)\psi_{2ij}-\sigma_{V\psi_2 j}\right)\right)$$

$$+o_p\left(N^{-2}\right)$$

$$=\frac{1}{N^2}\left(\sum_{p=1}^{M-1}\sum_{l=1}^{M-1}\sum_{t=1}^{M-1}\sum_{j=1}^{M-1}\omega_{pl}{}^{-1}\omega_{tj}{}^{-1}Cov\left(V_i-\tau,\psi_{2ip}\right)Cov\left(\psi_{2it}\psi_{2il},\left(V_i-\tau\right)\psi_{2ij}\right)\right)$$

$$+\frac{1}{N^2}\left(\sum_{p=1}^{M-1}\sum_{l=1}^{M-1}\sum_{t=1}^{M-1}\sum_{j=1}^{M-1}\omega_{pl}{}^{-1}\omega_{tj}{}^{-1}Cov\left(\psi_{2ip},\psi_{2it}\psi_{2il}\right)Cov\left(V_i-\tau,\left(V_i-\tau\right)\psi_{2ij}\right)\right)$$

$$+\frac{1}{N^2}\left(\sum_{p=1}^{M-1}\sum_{l=1}^{M-1}\sum_{t=1}^{M-1}\sum_{j=1}^{M-1}\omega_{pl}{}^{-1}\omega_{tj}{}^{-1}Cov\left(V_i-\tau,\psi_{2it}\psi_{2il}\right)Cov\left(\psi_{2ip},\left(V_i-\tau\right)\psi_{2ij}\right)\right)$$

$$+o_p\left(N^{-2}\right)$$

$$E\left(T_{N2}S_{N2}\right)$$

$$=-\frac{1}{N}\sigma_{V\psi_2}{}'\Omega^{-1}E\left(\left(\sum_{k=1}^{N}\begin{pmatrix}\psi_{2k1}\\ \vdots\\ \psi_{2k,M-1}\end{pmatrix}\right)\right.$$

$$\left.\times\frac{1}{N^3}\left(\sum_{p=1}^{M-1}\sum_{l=1}^{M-1}\sum_{t=1}^{M-1}\sum_{j=1}^{M-1}\sum_{r=1}^{N}\sum_{s=1}^{N}\sum_{i=1}^{N}\psi_{2rp}\omega_{pl}{}^{-1}\left(\psi_{2st}\psi_{2sl}-\omega_{tl}\right)\omega_{tj}{}^{-1}\left(\left(V_i-\tau\right)\psi_{2ij}-\sigma_{V\psi_2 j}\right)\right)\right)$$

$$=-\frac{1}{N^4}\sigma_{V\psi_2}{}'\Omega^{-1}E\begin{pmatrix}\displaystyle\sum_{p=1}^{M-1}\sum_{l=1}^{M-1}\sum_{t=1}^{M-1}\sum_{j=1}^{M-1}\sum_{k=1}^{N}\sum_{r=1}^{N}\sum_{s=1}^{N}\sum_{i=1}^{N}\omega_{pl}{}^{-1}\omega_{tj}{}^{-1}\psi_{2k1}\psi_{2rp}\left(\psi_{2st}\psi_{2sl}-\omega_{tl}\right)\left(\left(V_i-\tau\right)\psi_{2ij}-\sigma_{V\psi_2 j}\right)\\ \vdots\\ \displaystyle\sum_{p=1}^{M-1}\sum_{l=1}^{M-1}\sum_{t=1}^{M-1}\sum_{j=1}^{M-1}\sum_{k=1}^{N}\sum_{r=1}^{N}\sum_{s=1}^{N}\sum_{i=1}^{N}\omega_{pl}{}^{-1}\omega_{tj}{}^{-1}\psi_{2k,M-1}\psi_{2rp}\left(\psi_{2st}\psi_{2sl}-\omega_{tl}\right)\left(\left(V_i-\tau\right)\psi_{2ij}-\sigma_{V\psi_2 j}\right)\end{pmatrix}$$

For the same reason as mentioned above, we consider the three cases when $(k=r, s=i, r\neq i)$, $(k=i, r=s, s\neq i)$, and $(k=s, r=i, s\neq i)$. The equation above becomes

$$=-\frac{1}{N^4}\sigma_{V\psi_2}{}'\Omega^{-1}E\begin{pmatrix} \sum_{p=1}^{M-1}\sum_{l=1}^{M-1}\sum_{t=1}^{M-1}\sum_{j=1}^{M-1}\sum_{i=1}^{N}\sum_{r\neq i}\omega_{pl}{}^{-1}\omega_{tj}{}^{-1}\psi_{2r1}\psi_{2rp}\left(\psi_{2it}\psi_{2il}-\omega_{tl}\right)\left(\left(V_i-\tau\right)\psi_{2ij}-\sigma_{V\psi_2 j}\right) \\ \vdots \\ \sum_{p=1}^{M-1}\sum_{l=1}^{M-1}\sum_{t=1}^{M-1}\sum_{j=1}^{M-1}\sum_{i=1}^{N}\sum_{r\neq i}\omega_{pl}{}^{-1}\omega_{tj}{}^{-1}\psi_{2r,M-1}\psi_{2rp}\left(\psi_{2it}\psi_{2il}-\omega_{tl}\right)\left(\left(V_i-\tau\right)\psi_{2ij}-\sigma_{V\psi_2 j}\right) \end{pmatrix}$$

$$-\frac{1}{N^4}\sigma_{V\psi_2}{}'\Omega^{-1}E\begin{pmatrix} \sum_{p=1}^{M-1}\sum_{l=1}^{M-1}\sum_{t=1}^{M-1}\sum_{j=1}^{M-1}\sum_{i=1}^{N}\sum_{s\neq i}\omega_{pl}{}^{-1}\omega_{tj}{}^{-1}\psi_{2sp}\left(\psi_{2st}\psi_{2sl}-\omega_{tl}\right)\psi_{2i1}\left(\left(V_i-\tau\right)\psi_{2ij}-\sigma_{V\psi_2 j}\right) \\ \vdots \\ \sum_{p=1}^{M-1}\sum_{l=1}^{M-1}\sum_{t=1}^{M-1}\sum_{j=1}^{M-1}\sum_{i=1}^{N}\sum_{s\neq i}\omega_{pl}{}^{-1}\omega_{tj}{}^{-1}\psi_{2sp}\left(\psi_{2st}\psi_{2sl}-\omega_{tl}\right)\psi_{2i,M-1}\left(\left(V_i-\tau\right)\psi_{2ij}-\sigma_{V\psi_2 j}\right) \end{pmatrix}$$

$$-\frac{1}{N^4}\sigma_{V\psi_2}{}'\Omega^{-1}E\begin{pmatrix} \sum_{p=1}^{M-1}\sum_{l=1}^{M-1}\sum_{t=1}^{M-1}\sum_{j=1}^{M-1}\sum_{i=1}^{N}\sum_{s\neq i}\omega_{pl}{}^{-1}\omega_{tj}{}^{-1}\psi_{2s1}\left(\psi_{2st}\psi_{2sl}-\omega_{tl}\right)\psi_{2ip}\left(\left(V_i-\tau\right)\psi_{2ij}-\sigma_{V\psi_2 j}\right) \\ \vdots \\ \sum_{p=1}^{M-1}\sum_{l=1}^{M-1}\sum_{t=1}^{M-1}\sum_{j=1}^{M-1}\sum_{i=1}^{N}\sum_{s\neq i}\omega_{pl}{}^{-1}\omega_{tj}{}^{-1}\psi_{2s,M-1}\left(\psi_{2st}\psi_{2sl}-\omega_{tl}\right)\psi_{2ip}\left(\left(V_i-\tau\right)\psi_{2ij}-\sigma_{V\psi_2 j}\right) \end{pmatrix}$$

$$+o_p\left(N^{-2}\right)$$

$$=-\frac{1}{N^2}\sigma_{V\psi_2}{}'\Omega^{-1}\begin{pmatrix} \sum_{p=1}^{M-1}\sum_{l=1}^{M-1}\sum_{t=1}^{M-1}\sum_{j=1}^{M-1}\omega_{pl}{}^{-1}\omega_{tj}{}^{-1}Cov\left(\psi_{2i1},\psi_{2ip}\right)Cov\left(\psi_{2it}\psi_{2il},\left(V_i-\tau\right)\psi_{2ij}\right) \\ \vdots \\ \sum_{p=1}^{M-1}\sum_{l=1}^{M-1}\sum_{t=1}^{M-1}\sum_{j=1}^{M-1}\omega_{pl}{}^{-1}\omega_{tj}{}^{-1}Cov\left(\psi_{2i,M-1},\psi_{2ip}\right)Cov\left(\psi_{2it}\psi_{2il},\left(V_i-\tau\right)\psi_{2ij}\right) \end{pmatrix}$$

$$-\frac{1}{N^2}\sigma_{V\psi_2}{}'\Omega^{-1}\begin{pmatrix} \sum_{p=1}^{M-1}\sum_{l=1}^{M-1}\sum_{t=1}^{M-1}\sum_{j=1}^{M-1}\omega_{pl}{}^{-1}\omega_{tj}{}^{-1}Cov\left(\psi_{2ip},\psi_{2it}\psi_{2il}\right)Cov\left(\psi_{2i1},\left(V_i-\tau\right)\psi_{2ij}\right) \\ \vdots \\ \sum_{p=1}^{M-1}\sum_{l=1}^{M-1}\sum_{t=1}^{M-1}\sum_{j=1}^{M-1}\omega_{pl}{}^{-1}\omega_{tj}{}^{-1}Cov\left(\psi_{2ip},\psi_{2it}\psi_{2il}\right)Cov\left(\psi_{2i,M-1},\left(V_i-\tau\right)\psi_{2tj}\right) \end{pmatrix}$$

$$-\frac{1}{N^2}\sigma_{V\psi_2}{}'\Omega^{-1}\begin{pmatrix} \sum_{p=1}^{M-1}\sum_{l=1}^{M-1}\sum_{t=1}^{M-1}\sum_{j=1}^{M-1}\omega_{pl}{}^{-1}\omega_{tj}{}^{-1}Cov\left(\psi_{2s1},\psi_{2st}\psi_{2sl}\right)Cov\left(\psi_{2ip},\left(V_i-\tau\right)\psi_{2tj}\right) \\ \vdots \\ \sum_{p=1}^{M-1}\sum_{l=1}^{M-1}\sum_{t=1}^{M-1}\sum_{j=1}^{M-1}\omega_{pl}{}^{-1}\omega_{tj}{}^{-1}Cov\left(\psi_{2s,M-1},\psi_{2st}\psi_{2sl}\right)Cov\left(\psi_{2ip},\left(V_i-\tau\right)\psi_{2tj}\right) \end{pmatrix}$$

$$+o_p\left(N^{-2}\right)$$

$$E(T_{N1}S_{N3})$$

$$=\frac{1}{N}E\left(\left(\sum_{k=1}^{N}(V_k-\tau)\right)\cdot\frac{1}{N^3}\left(\sum_{t=1}^{M-1}\sum_{j=1}^{M-1}\sum_{r=1}^{N}\sum_{s=1}^{N}\sum_{i=1}^{N}\omega_{tj}{}^{-1}(V_r-\tau)\psi_{2st}\psi_{2ij}\right)\right)$$

$$=\frac{1}{N^4}E\left(\sum_{t=1}^{M-1}\sum_{j=1}^{M-1}\sum_{k=1}^{N}\sum_{r=1}^{N}\sum_{s=1}^{N}\sum_{i=1}^{N}\omega_{tj}{}^{-1}(V_k-\tau)(V_r-\tau)\psi_{2st}\psi_{2ij}\right)$$

For the same reason as mentioned above, we consider the three cases when $(k=r,s=i,r\neq i)$, $(k=i,r=s,s\neq i)$, and $(k=s,r=i,s\neq i)$. The equation above becomes

$$=\frac{1}{N^4}E\left(\sum_{t=1}^{M-1}\sum_{j=1}^{M-1}\sum_{i=1}^{N}\sum_{r\neq i}\omega_{tj}{}^{-1}(V_r-\tau)^2\psi_{2it}\psi_{2ij}\right)$$

$$+\frac{1}{N^4}E\left(\sum_{t=1}^{M-1}\sum_{j=1}^{M-1}\sum_{i=1}^{N}\sum_{s\neq i}\omega_{tj}{}^{-1}(V_i-\tau)\psi_{2ij}(V_s-\tau)\psi_{2st}\right)$$

$$+\frac{1}{N^4}E\left(\sum_{t=1}^{M-1}\sum_{j=1}^{M-1}\sum_{i=1}^{N}\sum_{s\neq i}\omega_{tj}^{-1}\left(V_s-\tau\right)\psi_{2st}\left(V_i-\tau\right)\psi_{2ij}\right)$$

$$+o_p\left(N^{-2}\right)$$

$$=\frac{1}{N^2}\left(\sum_{t=1}^{M-1}\sum_{j=1}^{M-1}\omega_{tj}^{-1}Var\left(V_i-\tau\right)Cov\left(\psi_{2it},\psi_{2ij}\right)\right)$$

$$+\frac{1}{N^2}\left(\sum_{t=1}^{M-1}\sum_{j=1}^{M-1}\omega_{tj}^{-1}Cov\left(\left(V_i-\tau\right)\psi_{2ij}\right)Cov\left(\left(V_i-\tau\right)\psi_{2it}\right)\right)$$

$$+\frac{1}{N^2}\left(\sum_{t=1}^{M-1}\sum_{j=1}^{M-1}\omega_{tj}^{-1}Cov\left(\left(V_i-\tau\right)\psi_{2it}\right)Cov\left(\left(V_i-\tau\right)\psi_{2ij}\right)\right)$$

$$+o_p\left(N^{-2}\right)$$

$$=\frac{1}{N^2}\left(\sum_{t=1}^{M-1}\sum_{j=1}^{M-1}\omega_{tj}^{-1}\sigma_V^{2}Cov\left(\psi_{2it},\psi_{2ij}\right)\right)$$

$$+\frac{2}{N^2}\left(\sum_{t=1}^{M-1}\sum_{j=1}^{M-1}\omega_{tj}^{-1}Cov\left(\left(V_i-\tau\right)\psi_{2ij}\right)Cov\left(\left(V_i-\tau\right)\psi_{2it}\right)\right)$$

$$+o_p\left(N^{-2}\right)$$

$$E\left(T_{N2}S_{N3}\right)$$

$$=-\frac{1}{N}\sigma_{V\psi_2}{}'\Omega^{-1}E\left(\sum_{k=1}^{N}\begin{pmatrix}\psi_{2k1}\\ \vdots\\ \psi_{2k,M-1}\end{pmatrix}\frac{1}{N^3}\left(\sum_{t=1}^{M-1}\sum_{j=1}^{M-1}\sum_{r=1}^{N}\sum_{s=1}^{N}\sum_{i=1}^{N}\omega_{tj}^{-1}\left(V_r-\tau\right)\psi_{2st}\psi_{2ij}\right)\right)$$

$$=-\frac{1}{N^4}\sigma_{V\psi_2}{}'\Omega^{-1}E\begin{pmatrix} \sum_{t=1}^{M-1}\sum_{j=1}^{M-1}\sum_{k=1}^{N}\sum_{r=1}^{N}\sum_{s=1}^{N}\sum_{i=1}^{N}\omega_{tj}{}^{-1}\left(V_r-\tau\right)\psi_{2k1}\psi_{2st}\psi_{2ij} \\ \vdots \\ \sum_{t=1}^{M-1}\sum_{j=1}^{M-1}\sum_{k=1}^{N}\sum_{r=1}^{N}\sum_{s=1}^{N}\sum_{i=1}^{N}\omega_{tj}{}^{-1}\left(V_r-\tau\right)\psi_{2k,M-1}\psi_{2st}\psi_{2ij} \end{pmatrix}$$

For the same reason as mentioned above, we consider the three cases when $(k=r, s=i, r\neq i)$, $(k=i, r=s, s\neq i)$, and $(k=s, r=i, s\neq i)$. The equation above becomes

$$=-\frac{1}{N^4}\sigma_{V\psi_2}{}'\Omega^{-1}E\begin{pmatrix} \sum_{t=1}^{M-1}\sum_{j=1}^{M-1}\sum_{i=1}^{N}\sum_{r\neq i}\omega_{tj}{}^{-1}\left(V_r-\tau\right)\psi_{2r1}\psi_{2it}\psi_{2ij} \\ \vdots \\ \sum_{t=1}^{M-1}\sum_{j=1}^{M-1}\sum_{i=1}^{N}\sum_{r\neq i}\omega_{tj}{}^{-1}\left(V_r-\tau\right)\psi_{2r,M-1}\psi_{2it}\psi_{2ij} \end{pmatrix}$$

$$-\frac{1}{N^4}\sigma_{V\psi_2}{}'\Omega^{-1}E\begin{pmatrix} \sum_{t=1}^{M-1}\sum_{j=1}^{M-1}\sum_{i=1}^{N}\sum_{s\neq i}\omega_{tj}{}^{-1}\psi_{2i1}\psi_{2ij}\left(V_s-\tau\right)\psi_{2st} \\ \vdots \\ \sum_{t=1}^{M-1}\sum_{j=1}^{M-1}\sum_{i=1}^{N}\sum_{s\neq i}\omega_{tj}{}^{-1}\psi_{2i,M-1}\psi_{2ij}\left(V_s-\tau\right)\psi_{2st} \end{pmatrix}$$

$$-\frac{1}{N^4}\sigma_{V\psi_2}{}'\Omega^{-1}E\begin{pmatrix} \sum_{t=1}^{M-1}\sum_{j=1}^{M-1}\sum_{i=1}^{N}\sum_{s\neq i}\omega_{tj}{}^{-1}\psi_{2s1}\psi_{2st}\left(V_i-\tau\right)\psi_{2ij} \\ \vdots \\ \sum_{t=1}^{M-1}\sum_{j=1}^{M-1}\sum_{i=1}^{N}\sum_{s\neq i}\omega_{tj}{}^{-1}\psi_{2s,M-1}\psi_{2st}\left(V_i-\tau\right)\psi_{2ij} \end{pmatrix}$$

$$+o_p\left(N^{-2}\right)$$

$$
=-\frac{1}{N^2}\sigma_{V\psi_2}{}'\Omega^{-1}\begin{pmatrix} \sum_{t=1}^{M-1}\sum_{j=1}^{M-1}\omega_{tj}{}^{-1}Cov\left(\left(V_r-\tau\right),\psi_{2r1}\right)Cov\left(\psi_{2it},\psi_{2ij}\right) \\ \vdots \\ \sum_{t=1}^{M-1}\sum_{j=1}^{M-1}\omega_{tj}{}^{-1}Cov\left(\left(V_r-\tau\right),\psi_{2r,M-1}\right)Cov\left(\psi_{2it},\psi_{2ij}\right) \end{pmatrix}
$$

$$
-\frac{1}{N^2}\sigma_{V\psi_2}{}'\Omega^{-1}\begin{pmatrix} \sum_{t=1}^{M-1}\sum_{j=1}^{M-1}\omega_{tj}{}^{-1}Cov\left(\psi_{2i1},\psi_{2ij}\right)Cov\left(\left(V_s-\tau\right),\psi_{2st}\right) \\ \vdots \\ \sum_{t=1}^{M-1}\sum_{j=1}^{M-1}\omega_{tj}{}^{-1}Cov\left(\psi_{2i,M-1},\psi_{2ij}\right)Cov\left(\left(V_s-\tau\right),\psi_{2st}\right) \end{pmatrix}
$$

$$
-\frac{1}{N^2}\sigma_{V\psi_2}{}'\Omega^{-1}\begin{pmatrix} \sum_{t=1}^{M-1}\sum_{j=1}^{M-1}\omega_{tj}{}^{-1}Cov\left(\psi_{2s1},\psi_{2st}\right)Cov\left(\left(V_i-\tau\right),\psi_{2ij}\right) \\ \vdots \\ \sum_{t=1}^{M-1}\sum_{j=1}^{M-1}\omega_{tj}{}^{-1}Cov\left(\psi_{2s,M-1},\psi_{2st}\right)Cov\left(\left(V_i-\tau\right),\psi_{2ij}\right) \end{pmatrix}
$$

$$
+o_p\left(N^{-2}\right)
$$

$$
E\left(T_{N1}S_{N4}\right)
$$

$$
=-E\left(\left(\frac{1}{n}\sum_k\left(V_k-\tau\right)\right)\left(\Omega^{-1}\left(\frac{1}{n}\sum_r\psi_{2r}\right)\right)'\sigma_{V\psi_2}\left(\Omega^{-1}\left(\frac{1}{n}\sum_s\psi_{2s}\right)\right)'\left(\frac{1}{n}\sum_i\psi_{2i}\right)\right)
$$

$$
=-E\left(\sigma_{V\psi_2}{}'\Omega^{-1}\left(\frac{1}{n}\sum_k\left(V_k-\tau\right)\right)\left(\frac{1}{n}\sum_r\psi_{2r}\right)\left(\frac{1}{n}\sum_s\psi_{2s}\right)'\Omega^{-1}\left(\frac{1}{n}\sum_i\psi_{2i}\right)\right)
$$

$$
=E\left(T_{N2}S_{N3}\right)
$$

$E\left(T_{N2}S_{N4}\right)$

$$=\frac{1}{N}\sigma_{V\psi_2}{}'\Omega^{-1}E\left(\sum_{k=1}^{N}\begin{pmatrix}\psi_{2k1}\\ \vdots\\ \psi_{2k,M-1}\end{pmatrix}\frac{1}{N^3}\left(\sum_{p=1}^{M-1}\sum_{l=1}^{M-1}\sum_{t=1}^{M-1}\sum_{j=1}^{M-1}\sum_{r=1}^{N}\sum_{s=1}^{N}\sum_{i=1}^{N}\sigma_{V\psi_2,p}\omega_{pl}{}^{-1}\psi_{2rl}\psi_{2st}\omega_{tj}{}^{-1}\psi_{2ij}\right)\right)$$

$$=\frac{1}{N^4}\sigma_{V\psi_2}{}'\Omega^{-1}E\begin{pmatrix}\sum_{p=1}^{M-1}\sum_{l=1}^{M-1}\sum_{t=1}^{M-1}\sum_{j=1}^{M-1}\sum_{k=1}^{N}\sum_{r=1}^{N}\sum_{s=1}^{N}\sum_{i=1}^{N}\sigma_{V\psi_2,p}\omega_{pl}{}^{-1}\omega_{tj}{}^{-1}\psi_{2k1}\psi_{2rl}\psi_{2st}\psi_{2ij}\\ \vdots\\ \sum_{p=1}^{M-1}\sum_{l=1}^{M-1}\sum_{t=1}^{M-1}\sum_{j=1}^{M-1}\sum_{k=1}^{N}\sum_{r=1}^{N}\sum_{s=1}^{N}\sum_{i=1}^{N}\sigma_{V\psi_2,p}\omega_{pl}{}^{-1}\omega_{tj}{}^{-1}\psi_{2k,M-1}\psi_{2rl}\psi_{2st}\psi_{2ij}\end{pmatrix}$$

For the same reason as mentioned above, we consider the three cases when $\left(k=r,s=i,r\neq i\right)$, $\left(k=i,r=s,s\neq i\right)$, and $\left(k=s,r=i,s\neq i\right)$. The equation above becomes

$$=\frac{1}{N^4}\sigma_{V\psi_2}{}'\Omega^{-1}E\begin{pmatrix}\sum_{p=1}^{M-1}\sum_{l=1}^{M-1}\sum_{t=1}^{M-1}\sum_{j=1}^{M-1}\sum_{i=1}^{N}\sum_{r\neq i}\sigma_{V\psi_2,p}\omega_{pl}{}^{-1}\omega_{tj}{}^{-1}\psi_{2r1}\psi_{2rl}\psi_{2it}\psi_{2ij}\\ \vdots\\ \sum_{p=1}^{M-1}\sum_{l=1}^{M-1}\sum_{t=1}^{M-1}\sum_{j=1}^{M-1}\sum_{i=1}^{N}\sum_{r\neq i}\sigma_{V\psi_2,p}\omega_{pl}{}^{-1}\omega_{tj}{}^{-1}\psi_{2r,M-1}\psi_{2rl}\psi_{2it}\psi_{2ij}\end{pmatrix}$$

$$+\frac{1}{N^4}\sigma_{V\psi_2}{}'\Omega^{-1}E\begin{pmatrix}\sum_{p=1}^{M-1}\sum_{l=1}^{M-1}\sum_{t=1}^{M-1}\sum_{j=1}^{M-1}\sum_{i=1}^{N}\sum_{s\neq i}\sigma_{V\psi_2,p}\omega_{pl}{}^{-1}\omega_{tj}{}^{-1}\psi_{2sl}\psi_{2st}\psi_{2i1}\psi_{2ij}\\ \vdots\\ \sum_{p=1}^{M-1}\sum_{l=1}^{M-1}\sum_{t=1}^{M-1}\sum_{j=1}^{M-1}\sum_{i=1}^{N}\sum_{s\neq i}\sigma_{V\psi_2,p}\omega_{pl}{}^{-1}\omega_{tj}{}^{-1}\psi_{2sl}\psi_{2st}\psi_{2i,M-1}\psi_{2ij}\end{pmatrix}$$

$$+\frac{1}{N^4}\sigma_{V\psi_2}{}'\Omega^{-1}E\begin{pmatrix} \sum_{p=1}^{M-1}\sum_{l=1}^{M-1}\sum_{t=1}^{M-1}\sum_{j=1}^{M-1}\sum_{i=1}^{N}\sum_{s\neq i}\sigma_{V\psi_2,p}\omega_{pl}{}^{-1}\omega_{tj}{}^{-1}\psi_{2s1}\psi_{2sl}\psi_{2it}\psi_{2ij} \\ \vdots \\ \sum_{p=1}^{M-1}\sum_{l=1}^{M-1}\sum_{t=1}^{M-1}\sum_{j=1}^{M-1}\sum_{i=1}^{N}\sum_{s\neq i}\sigma_{V\psi_2,p}\omega_{pl}{}^{-1}\omega_{tj}{}^{-1}\psi_{2s,M-1}\psi_{2sl}\psi_{2it}\psi_{2ij} \end{pmatrix}$$

$$+o_p\left(N^{-2}\right)$$

$$=\frac{1}{N^2}\sigma_{V\psi_2}{}'\Omega^{-1}\begin{pmatrix} \sum_{p=1}^{M-1}\sum_{l=1}^{M-1}\sum_{t=1}^{M-1}\sum_{j=1}^{M-1}\sigma_{V\psi_2,p}\omega_{pl}{}^{-1}\omega_{tj}{}^{-1}Cov\left(\psi_{2i1},\psi_{2il}\right)Cov\left(\psi_{2it},\psi_{2ij}\right) \\ \vdots \\ \sum_{p=1}^{M-1}\sum_{l=1}^{M-1}\sum_{t=1}^{M-1}\sum_{j=1}^{M-1}\sigma_{V\psi_2,p}\omega_{pl}{}^{-1}\omega_{tj}{}^{-1}Cov\left(\psi_{2i,M-1},\psi_{2il}\right)Cov\left(\psi_{2it},\psi_{2ij}\right) \end{pmatrix}$$

$$+\frac{1}{N^2}\sigma_{V\psi_2}{}'\Omega^{-1}\begin{pmatrix} \sum_{p=1}^{M-1}\sum_{l=1}^{M-1}\sum_{t=1}^{M-1}\sum_{j=1}^{M-1}\sigma_{V\psi_2,p}\omega_{pl}{}^{-1}\omega_{tj}{}^{-1}Cov\left(\psi_{2il},\psi_{2it}\right)Cov\left(\psi_{2i1},\psi_{2ij}\right) \\ \vdots \\ \sum_{p=1}^{M-1}\sum_{l=1}^{M-1}\sum_{t=1}^{M-1}\sum_{j=1}^{M-1}\sigma_{V\psi_2,p}\omega_{pl}{}^{-1}\omega_{tj}{}^{-1}Cov\left(\psi_{2il},\psi_{2it}\right)Cov\left(\psi_{2i,M-1},\psi_{2ij}\right) \end{pmatrix}$$

$$+\frac{1}{N^2}\sigma_{V\psi_2}{}'\Omega^{-1}\begin{pmatrix} \sum_{p=1}^{M-1}\sum_{l=1}^{M-1}\sum_{t=1}^{M-1}\sum_{j=1}^{M-1}\sigma_{V\psi_2,p}\omega_{pl}{}^{-1}\omega_{tj}{}^{-1}Cov\left(\psi_{2i1},\psi_{2it}\right)Cov\left(\psi_{2il},\psi_{2ij}\right) \\ \vdots \\ \sum_{p=1}^{M-1}\sum_{l=1}^{M-1}\sum_{t=1}^{M-1}\sum_{j=1}^{M-1}\sigma_{V\psi_2,p}\omega_{pl}{}^{-1}\omega_{tj}{}^{-1}Cov\left(\psi_{2i,M-1},\psi_{2it}\right)Cov\left(\psi_{2il},\psi_{2ij}\right) \end{pmatrix}$$

$$+o_p\left(N^{-2}\right)$$

Q.E.D.

PROOF OF LEMMA 1.2:

Follows from the Central Limit Theorem and the properties of consistency.

Q.E.D.

Proof of Theorem 1.2:

It can be seen that $E\left(T_{N2}^{\,2}\right)$, and $E\left(T_{N1}T_{N2}\right)$ are $O_p\left(N^{-1}\right)$. The remaining terms of $S(K)$ are $O_p\left(N^{-2}\right)$. $\hat{S}(K)$ equals $S(K)$ but composed of sample expectations.

Note that $S(K)$ is $O_p\left(N^{-1}\right)$ and $\hat{A}-A=O_p\left(N^{-1/2}\right) \Leftrightarrow \hat{A}\xrightarrow{p}A$ for any $\hat{A}$ and A. Given Lemma 2, it is readily observed that

$$\hat{E}\left(T_{N2}^{\,2}\right)-E\left(T_{N2}^{\,2}\right)=\frac{1}{N}O_p\left(N^{-1/2}\right)=O_p\left(N^{-3/2}\right) \tag{26}$$

$$\hat{E}\left(T_{N1}T_{N2}\right)-E\left(T_{N1}T_{N2}\right)=\frac{1}{N}O_p\left(N^{-1/2}\right)=O_p\left(N^{-3/2}\right). \tag{27}$$

This is because $E\left(T_{N2}^{\,2}\right)=\frac{C_1}{N}$ and $E\left(T_{N1}T_{N2}\right)=\frac{C_2}{N}$ where C_1 and C_2 are constants with respect to N. $\hat{C}_i-C_i=O_p\left(N^{-1/2}\right)$ for $i=\{1,2\}$ from Lemma 2. Also,

$$\hat{E}\left(A_{Ni}B_{Nj}\right)-E\left(A_{Ni}B_{Nj}\right)=\frac{1}{N^2}O_p\left(N^{-1/2}\right)=O_p\left(N^{-5/2}\right), \tag{28}$$

for (A, B, i, j) corresponding to the remaining terms composing $S(K)$. Equation (28) holds for the same reason as the previous two except, in this case, the constant (with respect to N) is being divided by N^2 as opposed to N.

Using the results of (26) – (28), it can now be shown that $\hat{S}(K)$ is higher order asymptotically optimal for selection of K with respect to $S(K)$. Donald and Newey (1999) show that a selection criterion is higher-order asymptotically optimal if

$$\sup_{K} \frac{\left|\hat{S}(K)-S(K)\right|}{S(K)}=o_p(1). \tag{29}$$

Note that

$$\sup_{K} \frac{\left|\hat{S}(K)-S(K)\right|}{S(K)}$$

$$=\sup_{K} \frac{\left|\hat{E}\left(T_{N2}^{2}\right)-E\left(T_{N2}^{2}\right)\right|}{S(K)}+\sup_{K} \frac{\left|\hat{E}(T_{N1}T_{N2})-E(T_{N1}T_{N2})\right|}{S(K)}$$

$$+\sup_{K} \frac{\left|\hat{E}\left(R_{N1}^{2}\right)-E\left(R_{N1}^{2}\right)\right|}{S(K)}+\sup_{K} \frac{\left|\hat{E}\left(R_{N2}^{2}\right)-E\left(R_{N2}^{2}\right)\right|}{S(K)}+\sup_{K} \frac{\left|\hat{E}(R_{N1}R_{N2})-E(R_{N1}R_{N2})\right|}{S(K)}$$

$$+\sup_{K} \frac{\left|\hat{E}(T_{N1}R_{N1})-E(T_{N1}R_{N1})\right|}{S(K)}+\sup_{K} \frac{\left|\hat{E}(T_{N2}R_{N1})-E(T_{N2}R_{N1})\right|}{S(K)}+\sup_{K} \frac{\left|\hat{E}(T_{N1}R_{N2})-E(T_{N1}R_{N2})\right|}{S(K)}$$

$$+\sup_{K} \frac{\left|\hat{E}(T_{N2}R_{N2})-E(T_{N2}R_{N2})\right|}{S(K)}+\sup_{K} \frac{\left|\hat{E}(T_{N1}S_{N1})-E(T_{N1}S_{N1})\right|}{S(K)}+\sup_{K} \frac{\left|\hat{E}(T_{N2}S_{N1})-E(T_{N2}S_{N1})\right|}{S(K)}$$

$$+\sup_{K} \frac{\left|\hat{E}(T_{N1}S_{N2})-E(T_{N1}S_{N2})\right|}{S(K)}+\sup_{K} \frac{\left|\hat{E}(T_{N2}S_{N2})-E(T_{N2}S_{N2})\right|}{S(K)}+\sup_{K} \frac{\left|\hat{E}(T_{N1}S_{N3})-E(T_{N1}S_{N3})\right|}{S(K)}$$

$$+\sup_{K}\frac{\left|\hat{E}\left(T_{N2}S_{N3}\right)-E\left(T_{N2}S_{N3}\right)\right|}{S(K)}+\sup_{K}\frac{\left|\hat{E}\left(T_{N1}S_{N4}\right)-E\left(T_{N1}S_{N4}\right)\right|}{S(K)}+\sup_{K}\frac{\left|\hat{E}\left(T_{N2}S_{N4}\right)-E\left(T_{N2}S_{N4}\right)\right|}{S(K)}.$$

The first two terms are $\frac{O_p\left(N^{-3/2}\right)}{O_p\left(N^{-1}\right)}=O_p\left(N^{-1/2}\right)$. Therefore they are also $o_p(1)$. The remaining terms are $\frac{O_p\left(N^{-5/2}\right)}{O_p\left(N^{-1}\right)}=O_p\left(N^{-3/2}\right)$. Therefore they are also $o_p(1)$. Hence, the entire sum is $o_p(1)$. Thus, $\hat{S}(K)$ satisfies (29) and is higher order asymptotically optimal for the choice of K with respect to the selection criterion $S(K)$ for the GMM model considered.

Q.E.D.

Proof of Theorem 2.1:

Let $\psi=\bar{\psi}$ and see the proofs above.

Q.E.D.

5. References

CHAMBERLAIN, G., (1987), "Asymptotic Efficiency in Estimation with Conditional Moment Restrictions," *Journal of Econometrics* 34, 305-334.

DONALD, S., G. IMBENS, AND W. NEWEY, (2001), "Empirical Likelihood Estimation and Consistent Tests with Conditional Moment Restrictions," *mimeo*.

DONALD, S. AND W. NEWEY, (1999), "Choosing the Number of Instruments," *mimeo*.

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

HANSON, L., J. HEATON, AND A. YARON, (1996), "Finite-Sample Properties of Some Alternative GMM Estimators," *Journal of Business and Economic Studies* 14 (3), 262-280.

HELLERSTEIN, J. AND G. IMBENS, (1999), "Imposing Moment Restrictions from Auxiliary Data by Weighting," *Review of Economics and Statistics*.

HIRANO, K., G. IMBENS, G. RIDDER, (2000), "Efficient Estimation of Average Treatment Effects Using the Estimated Propensity Score," *NBER Technical Working Paper 251*.

IMBENS, G., R. SPADY, AND P. JOHNSON, (1998), "Information Theoretic Approaches to Inference in Moment Condition Models," *Econometrica* 66 (2), 333-357.

LALONDE, R., (1986), "Evaluating the Econometric Evaluations of Training Programs with Experimental Data," *American Economic Review* 76 (4), 604-620.

LI, K., (1987), "Asymptotic Optimality for C_P, C_L, Cross-Validation and Generalized Cross-Validation: Discrete Index Set," *The Annals of Statistics* 15 (3), 958-975.

NEWEY, W. AND R. SMITH, (2000), "Asymptotic Bias and Equivalence of GEL and GMM Estimators," *mimeo*.

QIN, J. AND J. LAWLESS, (1994), "Generalized Estimating Equations," *Annals of Statistics* 22 (1), 300-325.

ROSENBAUM, P., (1987), "Model-Based Direct Adjustment," *Journal of the American Statistical Association* 82, 387-394.

ROSENBAUM, P. AND D. RUBIN, (1983), "The Central Role of the Propensity Score in Observational Studies for Causal Effects," *Biometrika* 70 (1), 41-55.