

Published in:
Missing Data Methods: Cross-sectional Methods and Applications
Advances in Econometrics, Volume 27A, 93-127
December 2011, Emerald Group Publishing Limited

Efficient Estimation of the Dose-Response Function under Ignorability using Subclassification on the Covariates*

Matias D. Cattaneo

Department of Economics, University of Michigan

Max H. Farrell

Department of Economics, University of Michigan

June 18, 2011

Abstract

This chapter studies the large sample properties of a subclassification-based estimator of the Dose-Response Function under Ignorability. Employing standard regularity conditions, it is shown that the estimator is root- n consistent, asymptotically linear, and semiparametric efficient in large samples. A consistent estimator of the standard-error is also developed under the same assumptions. In a Monte Carlo experiment we investigate the finite sample performance of this simple and intuitive estimator and compare it to others commonly employed in the literature.

Keywords: missing data, treatment effects, blocking, subclassification, stratification, semi-parametric efficiency.

*This chapter was prepared for the 9th annual Advances in Econometrics (AIE) conference on missing data methods. We thank an anonymous referee and the co-editor, David Drukker, for comments and suggestions.

1 INTRODUCTION

Treatment effect models are a prime example of a missing data problem. Units are assumed to have a collection of distinct random potential outcomes but, depending on their treatment status, only one of these outcomes is observed. The population parameters of interest in these models are usually some feature of the marginal distributions of the potential outcomes such as the means or quantiles. These parameters, however, are not identifiable from a random sample of observed outcomes and treatment statuses without further assumptions because of the potential presence of selectivity bias; a non-random missing data problem. A common identifying assumption in these models is called Ignorability, which includes a key restriction on the data generating process known as unconfoundedness or selection on observables. This assumption imposes random missing data after conditioning on a set of predetermined always-observed covariates, and permits the development of flexible inference procedures by first working conditionally on the covariates and then averaging out appropriately.

In the context of finite multi-valued treatment effects, a simple and interesting estimand is the Dose-Response Function (DRF), which describes the mean effect of each treatment level on the outcome of interest.¹ Under Ignorability, many different semiparametric estimators for the DRF may be constructed using flexible approaches, including nonparametric regression methods, matching techniques, inverse probability weighting schemes, procedures based on the estimated (generalized) propensity score, and hybrid procedures that combine some of these techniques.² These estimators, which include a preliminary nonparametric estimator, are well known to be root- n consistent (where n is the sample size) and asymptotically normal under appropriate regularity conditions, provided certain restrictions on the tuning and smoothing parameters involved in the estimation are satisfied. In most cases these estimators are also asymptotically linear and semiparametric efficient.

This chapter develops a new semiparametric efficient estimator of the DRF based on the idea of subclassification, blocking, or stratification on the observed predetermined covariates. The estimator proceeds by first dividing the support of the observed covariates into disjoint cells, also called blocks or strata, then carrying on inference using only observations within each cell, and finally averaging out appropriately. Intuitively, for cells “small enough,” the

¹See, e.g., Imbens (2000), Lechner (2001), Imai and van Dyk (2004), Cattaneo (2010), and references therein.

²For a review on the program evaluation and missing data literatures, see, e.g., Chen, Hong and Tarozzi (2004, 2008), Heckman and Vytlačil (2007), Imbens and Wooldridge (2009), Bang and Robins (2005), Tsiatis (2006), Wooldridge (2007), and references therein.

potential outcomes within each cell are approximately missing completely at random by virtue of Ignorability, leading to a consistent, asymptotically linear, and semiparametric efficient estimator under conventional regularity conditions. Moreover, using this idea we also develop a simple and intuitive consistent standard-error estimator, leading to asymptotically valid confidence intervals for the population parameter of interest.

The idea behind the semiparametric estimator discussed in this chapter may be traced back to the early work of Cochran (1968), who informally discusses the idea of subclassification with a univariate continuous covariate in observational studies. In this chapter we formally derive a first-order, asymptotically linear large sample approximation for a class of subclassification-based semiparametric estimators that allow for an arbitrary number of continuous covariates as well as an arbitrary large polynomial of approximation within each cell. These results are also connected to the work of Rosenbaum and Rubin (1983, 1984), who discuss inference by subclassifying observations based on the estimated propensity score in observational studies. In this chapter, however, subclassification is done directly on the observed covariates rather than on the estimated (generalized) propensity score, thereby avoiding preliminary nonparametric estimation of the propensity score and the related technical issues of generated regressors and random denominators. The ongoing work of Cattaneo, Imbens, Pinto, and Ridder (2009) addresses the delicate issue of subclassification-based inference using the estimated propensity score. The results in this chapter can be viewed as a first step toward developing the theoretical properties of such a procedure by considering the “known (generalized) propensity score” case, since a known propensity score may be treated as a univariate observed covariate and the results herein apply immediately.³

The subclassification-based estimator studied in this chapter may also be viewed as a two-step semiparametric estimator that depends on a special nonparametric procedure called Partitioning. In this chapter we exploit this idea, together with some of the recently developed asymptotic results presented in Cattaneo and Farrell (2011) for nonparametric partitioning estimators, to provide sufficient conditions for the efficient semiparametric estimation of the DRF, and to construct simple and easy-to-implement consistent standard-error estimators. We assess the performance of these large sample approximations in a Monte Carlo experiment.

The rest of the chapter is organized as follows. Section 2 introduces the multi-valued treatment effect model, discusses identification, and describes (both intuitively and formally)

³For further discussion on estimators combining subclassification and regression see, e.g., Imbens (2004) and Imbens and Wooldridge (2009).

the subclassification-based semiparametric estimator. Section 3 develops the asymptotic properties of this estimator, while Section 4 reports the main results of a simulation study. Finally, Section 5 summarizes the work presented here and discusses possible extensions. All technical derivations are contained in the Appendix.

2 MODEL, IDENTIFICATION AND ESTIMATOR

This chapter focuses on the estimation of the Dose-Response Function in the context of a (finite) multi-valued treatment effect model. Suppose that $(Y_i, X_i', T_i)'$, $i = 1, 2, \dots, n$, is an observed random sample, where Y_i is an outcome variable, $X_i \in \mathcal{X} \subset \mathbb{R}^d$ is a vector of continuous covariates, and $T_i \in \mathcal{T} = \{0, \dots, \tau\}$ with \mathcal{T} a finite set of treatments or groups. The procedure discussed below may be easily generalized to allow for discrete covariates by computing the estimator for each fixed distinct combination, and then averaging out appropriately, as it is standard in the literature. However, to simplify the discussion (and notation) we consider only continuous predetermined covariates. The outcome variable Y_i is assumed to satisfy $Y_i = D_{0,i}Y_i(0) + \dots + D_{\tau,i}Y_i(\tau)$, where $D_{t,i} = \mathbf{1}(T_i = t)$, $t = 0, \dots, \tau$, is a treatment or group indicator, and $Y_i(0), \dots, Y_i(\tau)$ are $\tau + 1$ random potential outcomes. ($\mathbf{1}(\cdot)$ denotes the indicator function.) For each unit $i = 1, \dots, n$, only one of the $\tau + 1$ potential outcomes is observed, according to the value of T_i . This leads to the fundamental problem of causal inference in the context of program evaluation (e.g., Holland (1986)), a classical missing data problem.

The estimand of interest is the DRF given by $\mu = (\mu_0, \dots, \mu_\tau)'$ with $\mu_t = \mathbb{E}[Y_i(t)]$. More general estimands are briefly discussed in Section 5, which summarizes potential extensions to the work undertaken in this chapter. Because all but one of the potential outcomes are missing for each unit, μ is not identifiable from the data without further assumptions. The following identification assumption is commonly used in the missing data and program evaluation literatures.

Assumption 1. (Weak Ignorability) For all $t \in \mathcal{T}$:

- (a) $Y_i(t) \perp\!\!\!\perp D_{t,i} \mid X_i$.
- (b) $0 < e_{\min} \leq \mathbb{P}[T_i = t \mid X_i]$.

Assumption 1(a) corresponds to a (weak) version of unconfoundedness or selection on observables, and implies that after conditioning on the observed covariates missing data occurs completely at random. This assumption is strong, but commonly employed in the literature.

Assumption 1(b) ensures that the generalized propensity score $e_t(x) = \mathbb{P}[T_i = t | X_i = x]$ is bounded away from zero, an important condition for semiparametric efficient estimation. This assumption, and different variations thereof, has been commonly used in the missing data, measurement error, and treatment effect literatures.

Assumption 1 implies that

$$\mu_t = \mathbb{E}[Y_i(t)] = \mathbb{E}\left[\frac{D_{t,i}Y_i}{e_t(X_i)}\right] = \mathbb{E}\left[\frac{\mathbb{E}[D_{t,i}Y_i|X_i]}{e_t(X_i)}\right] = \mathbb{E}[\mathbb{E}[Y_i|T_i = t, X_i]],$$

which leads to a variety of semiparametric plug-in (feasible) estimation approaches for the DRF. These alternative representations motivate inverse probability weighting, imputation, and projection estimation, among other possibilities. For a discussion of these alternative, well-known approaches see, e.g., Chen, Hong and Tarozzi (2004, 2008), Bang and Robins (2005), Imbens, Newey, and Ridder (2006), Tsiatis (2006), Heckman and Vytlacil (2007), Imbens and Wooldridge (2009), and references therein. Regardless of the particular identifying approach employed, in all cases at least one nonparametric estimator is required, unless the researcher is willing to impose strong parametric assumptions. Suitable implementations of flexible, semiparametric estimators are available in the literature when using local polynomials (including kernels) or sieves (including series), and these estimators are known to be asymptotically linear and semiparametric efficient under appropriate regularity conditions. (An important alternative estimator is the matching estimator of Abadie and Imbens (2006) which is not asymptotically linear.)

To motivate the subclassification estimator considered in this chapter, note that if the potential outcomes are assumed to be missing completely at random, that is, if $Y(t) \perp\!\!\!\perp D_t$, then a simple (possibly inefficient) estimator of μ_t is given by

$$\bar{Y}_t = \frac{1}{\bar{W}_t} \sum_{i=1}^n D_{t,i}Y_i, \quad \bar{W}_t = \sum_{i=1}^n D_{t,i},$$

which is a simple weighted average of the observed outcomes. However, if the data are not missing completely at random, \bar{Y}_t will be inconsistent for μ_t in general. Nonetheless, Assumption 1 leads to a similar idea based on subclassification on the observed covariates X_i . Suppose that \mathcal{X} is compact and that $\mathcal{P}_n = \{P_j : j = 1, \dots, J_n^d\}$ is a disjoint partition covering \mathcal{X} with typical cell P_j (implicit dependence on n through the partitioning scheme is suppressed for notational ease). Within each (small) cell P_j of the the partition \mathcal{P}_n , Assumption 1 implies that $Y(t)$ is “approximately” independent of D_t , suggesting the following

subclassification-based estimator:

$$\hat{\mu}_t = \sum_{j=1}^{J_n^d} \frac{N_j}{n} \bar{Y}_{j,t}, \quad \bar{Y}_{j,t} = \frac{1}{N_{j,t}} \sum_{i=1}^n \mathbf{1}_{P_j}(X_i) D_{t,i} Y_i,$$

$$N_j = \sum_{i=1}^n \mathbf{1}_{P_j}(X_i), \quad N_{j,t} = \sum_{i=1}^n \mathbf{1}_{P_j}(X_i) D_{t,i}, \quad \mathbf{1}_{P_j}(x) = \mathbf{1}(x \in P_j).$$

The “local” estimate $\bar{Y}_{j,t}$ is only well defined when $N_{j,t} > 0$, which is guaranteed in large samples by Assumption 1(b), provided that the cells are not too small. A proper definition of this estimator needs to account for the potential empty cells in finite samples, as done formally below. From an intuitive point of view $N_{j,t}/N_j \approx \mathbb{P}[D_{t,i} = 1 | X_i \in P_j] \approx e_t(X_i)$. Thus, under appropriate regularity conditions and for a fine enough partition, it is natural to expect that

$$\hat{\mu}_t = \frac{1}{n} \sum_{i=1}^n \sum_{j=1}^{J_n^d} \frac{N_j}{N_{j,t}} \mathbf{1}_{P_j}(X_i) D_{t,i} Y_i \approx \mu_t.$$

If all cells of the partition become small as $J_n^d \rightarrow \infty$, this subclassification-based estimator may be viewed as a semiparametric estimator given by

$$\hat{\mu}_t = \frac{1}{n} \sum_{i=1}^n \hat{\mu}_t(X_i), \quad \hat{\mu}_t(x) = \sum_{j=1}^{J_n^d} \frac{N_j}{N_{j,t}} \mathbf{1}_{P_j}(x) D_{t,i} Y_i,$$

where J_n^d corresponds to the tuning parameter underlying the nonparametric procedure. In fact, $\hat{\mu}_t(x)$ corresponds to a special case of the nonparametric estimator of a regression function known as Partitioning (see, e.g., Györfi, Kohler, Krzyżak, and Walk (2002, Chapter 4) and Cattaneo and Farrell (2011)).

Valid first-order, asymptotically linear, semiparametric inference requires a delicate choice of tuning and smoothing parameters so that the higher-order variance and the higher-order bias of the statistic are asymptotically negligible. For the partitioning estimator, J_n^d is the tuning parameter which “controls” the variance of the estimator: the smaller the cells (i.e., the larger J_n^d), the larger the variance. The bias, on the other hand, is (partially) determined by the “quality” of approximation: within each cell, the approximation is based on the sample mean of $D_{t,i} Y_i$, leading to an approximation error proportional to the inverse of the length of the cell. Thus, if bias is a concern, a natural way to improve the approximation is to use a more flexible polynomial in X_i within each block.

These insights lead to the following subclassification-based estimator, which is the main object of study in this chapter. The following notation is needed to formally describe the estimator. For fixed $K \in \mathbb{N}$, let $R(x)$ represent a column vector containing the complete polynomial basis of order $(K - 1)$ based on $x \in \mathbb{R}^d$, that is, for $x = (x_1, \dots, x_d)'$ and $\alpha = (\alpha_1, \dots, \alpha_d)' \in \mathbb{Z}_+^d$ (a multi-index), with $|\alpha| = \alpha_1 + \dots + \alpha_d$ and $x^\alpha = x_1^{\alpha_1} \dots x_d^{\alpha_d}$, each element of $R(x)$ is given by x^α for $\alpha \in \{\mathbf{a} \in \mathbb{Z}_+^d : |\mathbf{a}| \leq K - 1\}$. For example, if $d = 1$ then $R(x) = (1, x, x^2, \dots, x^{K-1})'$. Within each cell P_j , the basis is denoted by $R_j(x) = \mathbf{1}(x \in P_j)R(x)$. Using this notation, a subclassification-based estimator (of order $K - 1$) is given by

$$\hat{\mu}_t = \frac{1}{n} \sum_{i=1}^n \hat{\mu}_t(X_i), \quad \hat{\mu}_t(x) = \sum_{j=1}^{J_n^d} R_j(x)' \hat{\beta}_j, \quad \hat{\beta}_j = \mathbf{1}_{n,j} (R_{j,t}' R_{j,t})^{-1} R_{j,t}' Y,$$

$$R_{j,t} = (D_{t,1} R_j(X_1), \dots, D_{t,n} R_j(X_n))', \quad Y = (Y_1, \dots, Y_n)',$$

where $\mathbf{1}_{n,j} = \mathbf{1}(\lambda_{\min}(\hat{\Omega}_{j,t}) > c)$, with $\lambda_{\min}(A)$ the minimum eigenvalue of a matrix A , $\hat{\Omega}_{j,t} = R_{j,t}' R_{j,t} / (nq_j)$, $q_j = \mathbb{P}[X_i \in P_j]$, and c a fixed positive constant.

This estimator is quite intuitive: within each cell, the unknown regression function is approximated by a polynomial of order $K - 1$ in X_i , which is used to impute missing values for each observation, and then the imputed values are averaged out to obtain the final estimator. As shown in the Appendix, under appropriate regularity conditions, $\mathbf{1}_{n,j}$ takes the value 1 with probability approaching one, so that the least squares problem within each cell of the partition is (asymptotically) well defined.

3 LARGE SAMPLE RESULTS

This section describes the large sample properties of estimator introduced in the previous section. The following assumption imposes a set of simple restrictions on the data generating process.

Assumption 2. (a) X_i has compact support $\mathcal{X} \subset \mathbb{R}^d$, and its (Lebesgue) density is bounded and bounded away from zero.

(b) $\mathbb{E}[|Y_i(t)|^4 | X_i]$ is bounded for all $t \in \mathcal{T}$.

(c) $\mu_t(x)$ is S_μ -times continuously differentiable for all $t \in \mathcal{T}$.

(d) $e_t(x)$ is S_e -times continuously differentiable for all $t \in \mathcal{T}$.

Assumption 2(a) is important, and may be relaxed only when certain special partitioning schemes are employed and more stringent moment assumptions are imposed, but is otherwise difficult to weaken. Assumptions 2(b)-(d) implicitly control the rate of convergence in uniform norm of the partitioning nonparametric estimator, as shown in Cattaneo and Farrell (2011), and are standard in nonparametric and semiparametric estimation.

Regarding the partitioning nonparametric estimator, the following assumption will be imposed throughout. For scalars sequence $\{a_j : j = 1, \dots, J_n\}$, let $a_j \asymp J_n^{-1}$ denote that $C_* J_n^{-1} \leq a_j \leq C^* J_n^{-1}$ with C_* and C^* universal positive constants not depending on n nor $j = 1, \dots, J_n$.

Assumption 3. (a) For $\ell = 1, \dots, d$ and $J_n \in \mathbb{N}$, let the ℓ -dimension of \mathcal{X} be partitioned into the J_n disjoint intervals $[p_{\ell,j-1}, p_{\ell,j}]$, $j = 1, \dots, J_n - 1$, and $[p_{\ell,J_n-1}, p_{\ell,J_n}]$, satisfying $p_{\ell,j-1} < p_{\ell,j}$ for all j , and $|p_{\ell,j} - p_{\ell,j-1}| \asymp J_n^{-1}$. The complete partition of \mathcal{X} consists of the J_n^d sets formed as Cartesian products of all such intervals, with typical cell denoted P_j .

(b) For some $K \in \mathbb{N}$, $R(x)$ represents the complete polynomial basis of order $K - 1$ based on $x \in \mathbb{R}^d$.

Assumption 3(a) imposes natural restrictions on the partitioning scheme employed, which guarantee that each cell is well defined. By construction, each cell must satisfy: $\text{vol}(P_j) \asymp J_n^{-d}$, or equivalently, for some positive constants C_* and C^* : $C_* J_n^d \leq \min_{1 \leq j \leq J_n^d} \text{vol}(P_j) \leq \max_{1 \leq j \leq J_n^d} \text{vol}(P_j) \leq J_n^{-d} \leq C^* J_n^d$, where $\text{vol}(\cdot)$ denotes the volume of cell P_j . The simplest possible scheme is an evenly spaced partition, but Assumption 3(a) allows other possibilities so long as all cells continue to decrease proportionally to J_n^d . For a simple example, one may use a partition twice as fine in a region of abundant data compared to a sparse region (e.g., where the density is low). Assumption 3(b) specifies the degree of the polynomial used in the approximation within each cell. This assumption is meant to cover the general, unrestricted case, although in applications other (restricted) bases may be of interest. For example, if $\mu_t(x)$ is assumed to be additively separable, then the interactions between covariates may not be included in the basis $R(x)$, leading to a simpler least squares problem. The goal of Assumption 3(b) is to ensure that $R(x)$ is flexible enough to remove bias up to “order K ”, as shown in the Appendix.

The following theorem establishes that $\hat{\mu}_t$ has an asymptotically linear representation, with the well-known efficient influence function for μ_t (see, e.g., Hahn (1998)).

Theorem 1. *Suppose Assumptions 1–3 hold, and $\sqrt{n}J_n^{-K \wedge S_\mu \wedge S_e} \rightarrow 0$ and $J_n^{10d/7} \log(J_n)^2/n \rightarrow 0$. Then, for all $t \in \mathcal{T}$,*

$$\sqrt{n}(\hat{\mu}_t - \mu_t) = \frac{1}{\sqrt{n}} \sum_{i=1}^n \psi_t(Y_i, X_i, T_i) + o_p(1),$$

where

$$\psi_t(Y_i, X_i, T_i) = \frac{D_{t,i}(Y_i - \mu_t(X_i))}{e_t(X_i)} + \mu_t(X_i) - \mu_t.$$

Theorem 1 shows that there exists a choice of $J_n^d \rightarrow \infty$ such that $\hat{\mu}$ is asymptotically linear and semiparametric efficient, provided both $\mu_t(x)$ and $e_t(x)$ are smooth enough and K is large enough.⁴ The rate restrictions in Theorem 1 describe the lower and upper bounds on the rate of growth for the nonparametric tuning parameter, as is common in semiparametric inference. This condition formalizes the intuition above: the first statement requires sufficiently small cells to control bias, while the second ensures the nonparametric variance does not grow too fast.

It follows from this theorem that $\sqrt{n}(\hat{\mu} - \mu) \rightarrow_d \mathcal{N}(0, V)$, with V the semiparametric efficiency bound for μ , that is, V has (t, s) -element ($1 \leq t, s \leq \tau$) given by

$$V_{[t,s]} = \mathbb{E} \left[\mathbf{1}(t = s) \frac{\sigma_t^2(X_i)}{e_t(X_i)} + (\mu_t(X_i) - \mu_t)(\mu_s(X_i) - \mu_s) \right],$$

where $\sigma_t^2(X_i) = \mathbb{V}[Y_i(t)|X_i]$. See, e.g., Cattaneo (2010) for a discussion on this and related results.

In order to construct feasible, asymptotically valid confidence intervals for μ a consistent estimator of the standard errors is needed. Several alternatives are in principle possible, although a subclassification-based estimator seems most natural in the present context. One such estimator may be justified as follows. The overall asymptotic variance may be decomposed into the sum of the “within” and “between” variance as follows:

$$V_{[t,s]} = \mathbb{E} \left[\mathbf{1}(t = s) \frac{\sigma_t^2(X_i)}{e_t(X_i)} \right] + \mathbb{E} [(\mu_t(X_i) - \mu_t)(\mu_s(X_i) - \mu_s)] = V_{W,[t,s]} + V_{B,[t,s]}.$$

It is intuitive to separately estimate each component. First, because a least squares estimate is computed within each cell, a natural choice for $\hat{V}_{W,[t,s]}$ is a Huber-Eicker-White

⁴The rate restrictions imposed in Theorem 1 are in general not minimal, and may be relaxed in certain cases. It is possible to show by example that a necessary condition is given by $J_n^d/n \rightarrow 0$.

heteroskedasticity-robust estimator:

$$\hat{V}_{W,[t,s]} = \mathbf{1}(t = s) \sum_{j=1}^{J_n^d} \mathbf{1}_{n,j} \hat{L}'_j \hat{\Omega}_{j,t}^{-1} \hat{\Sigma}_{j,t} \hat{\Omega}_{j,t}^{-1} \hat{L}_j, \quad \hat{L}_j = \frac{1}{nq_j} \sum_{i=1}^n R_j(X_i),$$

$$\hat{\Sigma}_{j,t} = \frac{1}{n} \sum_{i=1}^n R_j(X_i) R_j(X_i)' D_{t,i} (Y_i - \hat{\mu}_t(X_i))^2.$$

This estimator has a simple, intuitive representation when $K = 1$, given by

$$\hat{V}_{W,[t,s]} = \mathbf{1}(t = s) \frac{1}{n} \sum_{i=1}^n \hat{\sigma}_t^2(X_i), \quad \hat{\sigma}_t^2(x) = \sum_{j=1}^{J_n^d} \frac{N_j^2}{N_{j,t}^2} \mathbf{1}_{P_j}(x) D_{t,i} (Y_i - \hat{\mu}_t(X_i))^2.$$

Second, for $V_{B,[t,s]}$, a simple partitioning-based plug-in estimator is:

$$\hat{V}_{B,[t,s]} = \frac{1}{n} \sum_{i=1}^n (\hat{\mu}_t(X_i) - \hat{\mu}_t)(\hat{\mu}_s(X_i) - \hat{\mu}_s).$$

The following theorem verifies that both estimators, $\hat{V}_{W,[t,s]}$ and $\hat{V}_{B,[t,s]}$, are indeed consistent for their population counterparts.

Theorem 2. *Suppose the conditions of Theorem 1 hold. Then, for all $t, s \in \mathcal{T}$, $\hat{V}_{W,[t,s]} \rightarrow_p V_{W,[t,s]}$ and $\hat{V}_{B,[t,s]} \rightarrow_p V_{B,[t,s]}$.*

It follows immediately from Theorem 2 that \hat{V} with typical (t, s) -element $(1 \leq t, s \leq \tau)$ given by $\hat{V}_{[t,s]} = \hat{V}_{W,[t,s]} + \hat{V}_{B,[t,s]}$ is a consistent estimator of V , leading to a consistent estimator of the semiparametric efficiency bound obtained in Theorem 1.

4 SIMULATIONS

In this section we report the results of a Monte Carlo study of the subclassification-based estimator. We focus on a binary treatment (i.e. $\tau = 2$) and conduct inference on the average treatment effect throughout, both for simplicity and to facilitate comparison with other results in the program evaluation literature.

The data generating process we consider is as follows. $X_{1i} \sim \text{Uniform}[-2, 2]$, $Y_i(0) = \mu_0 + X_{1i} + \eta_{0i}$, $Y_i(1) = \mu_1 + \exp\{X_{1i}\} - \mathbb{E}[\exp\{X_{1i}\}] + \eta_{1i}$, and $T_i = \mathbf{1}\{X_{1i}^3/3 - X_{1i} + \eta_{2i} > 0\}$, where the errors $\eta_{ki} \sim N(0, 2)$, $k = 0, 1, 2$, and are mutually independent. We also consider

a heteroskedastic variant of this model, in which $\eta_{1i} \sim N(0, 2X_i^2)$. Further, we extend these models to include a second covariate by generating $X_{2i} \sim \text{Uniform}[-2, 2]$ independently of X_{1i} then setting $Y_i(0) = \mu_0 + X_{1i} + X_{2i} + \eta_{0i}$, $Y_i(1) = \mu_1 + X_{1i}^3 X_{2i}^2 + \exp\{X_{1i}\} + \exp\{X_{2i}\} - 2\mathbb{E}[\exp\{X_{1i}\}] + \eta_{1i}$, and all else as above. In all cases we set $\mu_0 = 0$ and $\mu_1 = 1$, so that the average treatment effect is one. We conduct simulations of each model with sample sizes of 500 and 1,000, both using 2,000 replications and evenly spaced cells.

The average bias for the univariate model are reported in Figure 1 for a range of J_n . The homoskedastic and heteroskedastic model produce very similar results, and the discussion below applies to both. The figure demonstrates several salient features of the subclassification estimator. First, the increased flexibility of the nonparametric estimator resultant from increasing the number of cells initially decreases the bias. The nonparametric procedure relies on $J_n^d \rightarrow \infty$ as $n \rightarrow \infty$, and the bias decreases accordingly: see (A-2) in the appendix. The second important feature is the choice of the (fixed) parameter K , giving the order of the fit within each cell. Recall from above that $K = 1$ corresponds to fitting means within each cell, $K = 2$ gives a linear fit, and so forth. As aforementioned, a larger K improves the theoretic bias properties of the estimator and for modest values of J_n this is borne out in Figure 1: For J_n below 10, fitting means within each cell may not be sufficient to remove bias, but an increase merely to linear fits is often adequate. Much more modest improvements result from a quadratic fit.

However, as J_n increases further, the bias properties decline: the estimator has increased bias compared to fewer cells, and substantially so for the quadratic fit. This is a consequence of the least squares problem being ill-posed in an increasing number of cells. Heuristically, for the (fixed) n chosen, these J_n represent sequences for which the rate restrictions of Theorem 1 do not hold, and hence the distributional approximation is invalid. Recalling the formulation above, for these “empty cells” $\mathbf{1}_{n,j} = 0$ and the matrix $\hat{\Omega}_{j,t}$ is singular (or near singular; in practice a numerical cut-off is employed). Hence, these cells are not included in the estimate, leading to bias. It is beyond the scope of this work to study a formal trimming procedure, but one that controls for empty cells in a systematic way may lead to improved performance for certain choices of J_n . These results (and similarly those of the bivariate specification below) may be interpreted as a cautionary tale regarding choice of smoothing and tuning parameters in nonparametric estimation in general. It is also important to note that this phenomenon does not impact the estimator with degree zero (fitting means) as severely, since only one observation per cell is required. Indeed, for bias the piecewise constant version of the subclassification-based estimator is quite robust to the choice of J_n . Finally, note that

the increased sample size expands the range of J_n for which the estimator performs well, for any choice of K .

Figure 2 reports coverage rates for 95% confidence intervals for the univariate models. Many of the same conclusions are evident. For modest values of J_n , increased K leads to more accurate coverage, but beyond a certain value, coverage declines more rapidly for higher values of K . Again, the robustness to choice of J_n for $K = 1$ is evident. The coverage is remarkably accurate for even a large number of cells. The variance estimator accounts for heteroskedasticity quite well: only a small loss is evident. In practice, the “empty-cell” issue is likely to be a greater concern.

Figures 3–6 show the Gaussian approximation for the four univariate models. The estimator approximates the semiparametric efficiency bound for several different choices of J_n and K , matching the result of Theorem 1. In all cases, the same conclusions above are evident and the heteroskedasticity makes little difference. For moderate choices, the robustness is again demonstrated. However, for $n = 500$ and a large J_n , the estimator is biased and the variance is inflated: the bottom-right graph in Figures 3 and 5 shows that the approximation can be poor. Again, increasing the sample size ameliorates the issue, as would be expected from the theory in Section 3.

The conclusions from the bivariate model are somewhat different. Figures 7 and 8 report bias and coverage for the bivariate models. Note that the range of J_n is restricted compared to the univariate models. Recall that the theory requires J_n^d cells, so that the points marked as “10” in Figures 7 and 8 actually utilize $J_n^2 = 100$ total cells. Here the empty cell problem has become severe, and both the bias and coverage properties become extremely poor. Also observe that for smaller values of J_n , fitting means is no longer sufficient to remove bias or produce accurate coverage: both are more accurate for the quadratic fit for a larger range of J_n . This illustrates the tension between the bias and variance conditions in Theorem 1 for the sequence J_n . This “curse of dimensionality” is a common problem in nonparametric estimation. Figures 9–12 show the Gaussian approximations, which exhibit the same issues. In some cases, the approximation is extremely poor for these sample sizes. However, observe that for moderate values of J_n (e.g., $J_n = 3$, $J_n^d = 9$), the semiparametric efficiency bound is approximated well for certain choices of K . To investigate this further, we simulated the bivariate homoskedastic model with a sample size of $n = 2,000$. The Gaussian approximation is shown in Figure 13. As would be expected, the estimator performs better for a wider range of J_n and K . The bias and coverage results (not shown) are also substantially improved. When considering additional regressors, researchers should keep this “curse of

dimensionality” in mind.

Finally, we compare the partitioning estimator to several others common in the literature: inverse probability weighting (IPW), a series-based imputation estimator, and M-nearest-neighbor matching. The propensity score is estimated using a logistic regression on a power series of X_i up to order four or six, and then the average treatment effect is estimated by inverse weighting as in Hirano, Imbens, and Ridder (2003). The series estimator uses non-parametric regression to impute missing outcomes in much the same spirit as the partitioning estimator, but the approximation is global, see Imbens, Newey, and Ridder (2006). Here we use a power series of degree four or six, but to approximate the underlying regression function instead of the propensity score. Finally, we consider nearest-neighbor matching Abadie and Imbens (2006). We implement this in Stata using the `nnmatch` command of Abadie, Drukker, Herr, and Imbens (2004). We consider one- and two-neighbor matching, as well as simple and bias-adjusted estimates. For brevity, we consider only the univariate homoskedastic model. Following the above results, we use only 7- and 10-cell partitions, and only up to a linear fit. Table 1 presents mean-square error comparison between the estimators. Gaussian approximations are given in Figures 14 and 15. In the figures the 10-cell subclassification estimator with degree zero is given by the solid line, with the long-dashed line for degree one. Results are comparable with $J_n = 7$, so this is excluded. The comparison estimator is given by the short-dashed line for the “lower” degree (power series of degree four in the case of IPW and Series, or one match) and a dotted line for the “higher” degree (degree six, two matches).

The subclassification estimator performs comparably to these alternatives. Both the IPW and series estimators are known to attain the semiparametric efficiency bound, which is borne out in panels (A) and (B) of Figures 14 and 15. Table 1 shows that these estimators exhibit comparable variance to the subclassification estimator. For a fixed number of matches, nearest-neighbor matching is well-centered but does not attain the bound, and hence it is not surprising that the subclassification estimator is more concentrated, see panels (C) and (D). The MSE of the matching estimator is larger as a result. For a piecewise constant or linear fit, the subclassification estimator appears to be on par with popular choices in the econometrics literature.

5 EXTENSIONS AND FINAL REMARKS

The main result of this chapter (Theorem 1) shows that the subclassification-based estimator of the Dose-Response Function introduced in Section 2 is asymptotically linear and semiparametric efficient under standard regularity conditions. Theorem 2 also demonstrates that a simple, consistent standard errors estimator is easy to construct based on the idea of subclassification. In addition, the simulation study reported in Section 4 suggests that this estimator performs well in finite samples.

The theoretical results presented in this chapter may be easily extended to cover other potential estimands of interest. Perhaps the most natural extension would be to consider estimating the quantiles of the distribution of $Y(t)$, $t \in \mathcal{T}$. (See, e.g., Firpo (2007).) In this case, because the α -th quantile of $Y(t)$, denoted by $q_t(\alpha)$, solves $0 = \mathbb{E}[m(Y(t), q_t(\alpha); \alpha)]$ with $m(y, q; \alpha) = \mathbf{1}(y \leq q) - \alpha$, a natural subclassification-based estimator of $q_t(\alpha)$ would be given by $\hat{q}_t(\alpha) = \arg \min_q |M_n(q; \alpha)|$,

$$M_n(q; \alpha) = \frac{1}{n} \sum_{i=1}^n \hat{q}_t(X_i; \alpha), \quad \hat{q}_t(x; \alpha) = \sum_{j=1}^{J_n^d} R_j(X_i)' \hat{\beta}_{j,\alpha},$$

$$\hat{\beta}_{j,\alpha} = \mathbf{1}_{n,j} (R'_{j,t} R_{j,t})^{-1} R'_{j,t} Y(q; \alpha), \quad Y(q; \alpha) = (m(Y_1, q; \alpha), \dots, m(Y_n, q; \alpha))'.$$

Under appropriate regularity conditions, it seems plausible that the resulting estimator $\hat{q}_t(\alpha)$ would also be asymptotically normal and semiparametric efficient. More generally, it is natural to expect that such a result would hold for other estimands as defined by a choice of function $m(\cdot)$ in some appropriate class. For a discussion on related ideas and other potential extensions see, e.g., Cattaneo (2010) and references therein. These extensions are not considered in this chapter for brevity, and consequently are relegated for future work.

Another useful extension to the present work would be to develop a guide for the choice of J_n in applications. The number of cells is the nonparametric tuning parameter, and its choice is important for the finite sample properties of the estimator, as discussed in Section 4. A natural criterion for choosing J_n would be to consider a mean-square error expansion of the estimator, which could be minimized to find the optimal number of cells. Among other things, this would be a function of K , the smoothing parameter. Following this, a simple “plug-in” estimate could be proposed for the optimal J_n .

A APPENDIX

Throughout the appendix, C denotes a generic positive constant that may take different values in different places. All bounds are uniform in $j = 1, \dots, J_n^d$ unless explicitly noted otherwise. For A , a scalar, vector, or matrix, let $|A|$ denote the Euclidean norm.

Define $\Omega_{j,t} = q_j^{-1} \mathbb{E}[\mathbf{1}_{P_j}(X_i) D_{t,i} R(X_i) R(X_i)']$, $\varepsilon_t = (Y_1(t) - \mu_t(X_1), \dots, Y_n(t) - \mu_t(X_n))'$, and $E_t = ((e_t(X_1) - D_{t,1})/e_t(X_1), \dots, (e_t(X_n) - D_{t,n})/e_t(X_n))'$. We now collect several useful results regarding the nonparametric partition regression estimator. Details and proofs may be found in Cattaneo and Farrell (2011). All results given in the appendix implicitly utilize an appropriate non-singular linear transformation of the polynomial basis, although the same notation is maintained for simplicity. Cattaneo and Farrell (2011) give details on the appropriate rotation and demonstrate its existence under the conditions imposed in Theorem 1.

Lemma 1. *Under the conditions of Theorem 1, the following results hold:*

$$(A-1) \max_{1 \leq j \leq J_n^d} \sup_{x \in P_j} |R_j(x)| \leq C < \infty.$$

(A-2) *There exists vectors $\gamma_{\mu,j}$ and $\gamma_{e,j}$, $j = 1, \dots, J_n^d$, such that*

$$\max_{1 \leq j \leq J_n^d} \sup_{x \in P_j} |\mu_t(x) - R_j(x)' \gamma_{\mu,j}| = O(J_n^{-K \wedge S_\mu}),$$

and

$$\max_{1 \leq j \leq J_n^d} \sup_{x \in P_j} \left| \frac{1}{e_t(x)} - R_j(x)' \gamma_{e,j} \right| = O(J_n^{-K \wedge S_e}).$$

$$(A-3) \lambda_{\min}(\Omega_{j,t}) \geq C > 0.$$

$$(A-4) \max_{1 \leq j \leq J_n^d} \left| \hat{\Omega}_{j,t} - \Omega_{j,t} \right|^2 = O_p(J_n^d \log(J_n)/n).$$

$$(A-5) \max_{1 \leq j \leq J_n^d} \left| R_{j,t}' \varepsilon_t / (nq_j) \right|^2 = O_p(J_n^{9d/7} \log(J_n)^{5/7}/n).$$

$$(A-6) \max_{1 \leq j \leq J_n^d} \left| R_j' E_t / (nq_j) \right|^2 = O_p(J_n^d \log(J_n)/n).$$

$$(A-7) \max_{1 \leq j \leq J_n^d} \sup_{x \in P_j} |\hat{\mu}_{j,t}(x) - \mu_{j,t}(x)|^2 = O_p(J_n^{9d/7} \log(J_n)^{5/7}/n + J_n^{-2K \wedge S_\mu}).$$

Results (A-3) and (A-4) imply that $\max_{1 \leq j \leq J_n^d} |\Omega_{j,t}^{-1}| \leq C$, $\max_{1 \leq j \leq J_n^d} |\hat{\Omega}_{j,t}^{-1}| = O_p(1)$, and $\mathbb{P}(\max_{1 \leq j \leq J_n^d} |\mathbf{1}_{n,j} - 1| = 0) \rightarrow 1$.

A.1 PROOF OF THEOREM 1

Let $\gamma_{\mu,j}$ and $\gamma_{e,j}$ be as given in (A-4). Observe that

$$\sqrt{n}(\hat{\mu}_t - \mu_t) = \frac{1}{\sqrt{n}} \sum_{i=1}^n \psi_t(Y_i, X_i, T_i) + \epsilon_{n,1} + \epsilon_{n,2} + \epsilon_{n,3} + \epsilon_{n,4} + \epsilon_{n,5} + \epsilon_{n,6},$$

where

$$\begin{aligned} \epsilon_{n,1} &= \frac{1}{\sqrt{n}} \sum_{j=1}^{J_n^d} \sum_{i=1}^n \left(1 - \frac{D_{t,i}}{e_t(X_i)}\right) \mathbf{1}_{n,j} R_j(X_i)' (R'_{j,t} R_{j,t})^{-1} R'_{j,t} \varepsilon_t, \\ \epsilon_{n,2} &= \frac{1}{\sqrt{n}} \sum_{j=1}^{J_n^d} \sum_{i=1}^n \mathbf{1}_{n,j} \left(\frac{1}{e_t(X_i)} - \gamma'_{e,j} R_j(X_i)\right) D_{t,i} R_j(X_i)' (R'_{j,t} R_{j,t})^{-1} R'_{j,t} \varepsilon_t, \\ \epsilon_{n,3} &= \frac{1}{\sqrt{n}} \sum_{j=1}^{J_n^d} \sum_{i=1}^n \mathbf{1}_{n,j} \left(\gamma'_{e,j} R_j(X_i) - \frac{1}{e_t(X_i)}\right) \mathbf{1}_{P_j}(X_i) D_{t,i} (Y_i - \mu_t(X_i)), \\ \epsilon_{n,4} &= \frac{1}{\sqrt{n}} \sum_{j=1}^{J_n^d} \sum_{i=1}^n \sum_{k=1}^n \mathbf{1}_{n,j} R_j(X_i)' (R'_{j,t} R_{j,t})^{-1} D_{t,k} R_j(X_k) (\mu_t(X_k) - R_j(X_k)' \gamma_{\mu,j}), \\ \epsilon_{n,5} &= \frac{1}{\sqrt{n}} \sum_{j=1}^{J_n^d} \sum_{i=1}^n \mathbf{1}_{n,j} \mathbf{1}_{P_j}(X_i) (R(X_i)' \gamma_{\mu,j} - \mu_t(X_i)), \\ \epsilon_{n,6} &= \frac{1}{\sqrt{n}} \sum_{j=1}^{J_n^d} \sum_{i=1}^n (\mathbf{1}_{n,j} - 1) \mathbf{1}_{P_j}(X_i) \left\{ \frac{D_{t,i} (Y_i - \mu_t(X_i))}{e_t(X_i)} + \mu_t(X_i) \right\}. \end{aligned}$$

Consider each reminder $\epsilon_{n,1}-\epsilon_{n,6}$. First, $\epsilon_{n,1} = \epsilon_{n,11} + \epsilon_{n,12} + \epsilon_{n,13}$ with

$$\begin{aligned} \epsilon_{n,11} &= \frac{1}{\sqrt{n}} \sum_{j=1}^{J_n^d} \mathbf{1}_{n,j} [R'_j E_t]' \Omega_{j,t}^{-1} [\Omega_{j,t} - \hat{\Omega}_{j,t}] \Omega_{j,t}^{-1} [\Omega_{j,t} - \hat{\Omega}_{j,t}] \hat{\Omega}_{j,t}^{-1} [R'_{j,t} \varepsilon_t / (nq_j)] = o_p(1), \\ \epsilon_{n,12} &= -\frac{1}{\sqrt{n}} \sum_{j=1}^{J_n^d} \mathbf{1}_{n,j} [R'_j E_t]' \Omega_{j,t}^{-1} [\hat{\Omega}_{j,t} - \Omega_{j,t}] \Omega_{j,t}^{-1} [R'_{j,t} \varepsilon_t / (nq_j)] = o_p(1), \\ \epsilon_{n,13} &= \frac{1}{\sqrt{n}} \sum_{j=1}^{J_n^d} \mathbf{1}_{n,j} [R'_j E_t]' \Omega_{j,t}^{-1} [R'_{j,t} \varepsilon_t / (nq_j)] = o_p(1), \end{aligned}$$

because

$$\begin{aligned} |\epsilon_{n,11}| &\leq \sqrt{n} \max_{1 \leq j \leq J_n^d} |R'_j E_t / (nq_j)| \max_{1 \leq j \leq J_n^d} |\mathbf{1}_{n,j} \hat{\Omega}_{j,t}^{-1}| \max_{1 \leq j \leq J_n^d} |\hat{\Omega}_{j,t} - \Omega_{j,t}|^2 \max_{1 \leq j \leq J_n^d} |\Omega_{j,t}^{-1}|^2 \max_{1 \leq j \leq J_n^d} |R'_{j,t} \varepsilon_t / (nq_j)| \\ &= \sqrt{n} O_p(J_n^{3d/2} \log(J_n)^{3/2} / n^{3/2}) O_p(J_n^{9d/14} \log(J_n)^{5/14} / \sqrt{n}) = o_p(1), \end{aligned}$$

and simple variance bounds give $\mathbb{E}[\epsilon_{n,12}^2] = O(J_n^{2d}/n^2) = o_p(1)$ and $\mathbb{E}[\epsilon_{n,13}^2] = O(J_n^d/n) = o_p(1)$, as $\mathbb{E}[R_j(X_i)E_{t,i}|X_i] = 0$, $\mathbb{E}[q_j^{-1}D_{t,i}R_j(X_i)R_j(X_i)' - \Omega_{t,j}] = 0$ and $\mathbb{E}[D_{t,i}R_j(X_i)\varepsilon_{t,i}|X_i, T_i] = 0$.

Next, observe that $\epsilon_{n,2} = \epsilon_{n,21} + \epsilon_{n,22}$ with

$$\epsilon_{n,21} = -\frac{1}{\sqrt{n}} \sum_{j=1}^{J_n^d} \mathbf{1}_{n,j} \sum_{i=1}^n \left(\frac{1}{e_t(X_i)} - \gamma'_{e,j} R_j(X_i) \right) D_{t,i} R_j(X_i)' \hat{\Omega}_{j,t}^{-1} [\hat{\Omega}_{j,t} - \Omega_{j,t}] \Omega_{j,t}^{-1} [R'_{j,t} \varepsilon_t / (nq_j)] = o_p(1),$$

$$\epsilon_{n,22} = \frac{1}{\sqrt{n}} \sum_{j=1}^{J_n^d} \mathbf{1}_{n,j} \sum_{i=1}^n \left(\frac{1}{e_t(X_i)} - \gamma'_{e,j} R_j(X_i) \right) D_{t,i} R_j(X_i)' \Omega_{j,t}^{-1} [R'_{j,t} \varepsilon_t / (nq_j)] = o_p(1),$$

because

$$\begin{aligned} |\epsilon_{n,21}| &\leq O(J_n^{-K \wedge S_e}) \max_{1 \leq j \leq J_n^d} |\mathbf{1}_{n,j} \hat{\Omega}_{j,t}^{-1}| \max_{1 \leq j \leq J_n^d} |\hat{\Omega}_{j,t} - \Omega_{j,t}| \\ &\quad \times \max_{1 \leq j \leq J_n^d} |\Omega_{j,t}^{-1}| \max_{1 \leq j \leq J_n^d} |R'_{j,t} \varepsilon_t / (nq_j)| \frac{1}{\sqrt{n}} \sum_{j=1}^{J_n^d} \sum_{i=1}^n \mathbf{1}_{P_j}(X_i) \\ &= O(J_n^{-K \wedge S_e}) \sqrt{n} O_p(J_n^{d/2} \log(J_n)^{1/2} / \sqrt{n}) O_p(J_n^{9d/14} \log(J_n)^{5/14} / \sqrt{n}) = o_p(1), \end{aligned}$$

and $\mathbb{E}[\epsilon_{n,22}^2] = O(J_n^{-2K \wedge S_e}) = o_p(1)$.

Next, $\epsilon_{n,3} = o_p(1)$ because

$$\mathbb{E}[\epsilon_{n,3}^2] \leq \sum_{j=1}^{J_n^d} \mathbb{E} \left[\mathbf{1}_{P_j}(X_i) \left(\gamma'_{e,j} R(X_i) - \frac{1}{e_t(X_i)} \right)^2 (Y_i(t) - \mu_t(X_i))^2 \right] = O(J_n^{-2K \wedge S_e}) = o(1).$$

Next, $\epsilon_{n,4} = o_p(1)$ because

$$\begin{aligned} |\epsilon_{n,4}| &\leq \sqrt{n} O(J_n^{-K \wedge S_\mu}) \frac{J_n^d}{n^2} \sum_{j=1}^{J_n^d} \sum_{i=1}^n \sum_{k=1}^n \mathbf{1}_{n,j} R_j(X_i)' \hat{\Omega}_{j,t}^{-1} D_{t,k} R_j(X_k) \\ &\leq \sqrt{n} O_p(J_n^{-K \wedge S_\mu}) \frac{J_n^d}{n^2} \sum_{j=1}^{J_n^d} \sum_{i=1}^n \sum_{k=1}^n \mathbf{1}_{P_j}(X_i) \mathbf{1}_{P_j}(X_k) = \sqrt{n} O_p(J_n^{-K \wedge S_\mu}) = o_p(1). \end{aligned}$$

Next, $\epsilon_{n,5} = o_p(1)$ because

$$|\epsilon_{n,5}| \leq \sqrt{n}O(J_n^{-K \wedge S_\mu}) \frac{1}{n} \sum_{j=1}^{J_n^d} \sum_{i=1}^n \mathbf{1}_{P_j}(X_i) = \sqrt{n}O(J_n^{-K \wedge S_\mu}).$$

Finally, $\epsilon_{n,6} = o_p(1)$ because $\mathbb{P}(\max_{1 \leq j \leq J_n^d} |\mathbf{1}_{n,j} - 1| = 0) \rightarrow 1$. \blacksquare

A.2 PROOF OF THEOREM 2

For $\hat{V}_{W,[t,s]}$, first define $\tilde{\Sigma}_{j,t} = n^{-1} \sum_{i=1}^n R_j(X_i)R_j(X_i)'D_{t,i}(Y_i - \mu_t(X_i))^2$ and $\tilde{L}_j = \frac{1}{nq_j} \sum_{i=1}^n R_j(X_i)D_{t,i}/e_t(X_i)$, then note that

$$\hat{V}_{W,[t,t]} = \frac{1}{n} \sum_{i=1}^n \mathbf{1}_{n,j} \frac{D_{t,i}\varepsilon_{t,i}^2}{e_t(X_i)} + \epsilon_{W,n,1} + \epsilon_{W,n,2} + \epsilon_{W,n,3} + \epsilon_{W,n,4} + \epsilon_{W,n,5},$$

where

$$\begin{aligned} \epsilon_{W,n,1} &= \sum_{j=1}^{J_n^d} \mathbf{1}_{n,j} \hat{L}_j' \hat{\Omega}_{tj}^{-1} (\hat{\Sigma}_{tj} - \tilde{\Sigma}_{tj}) \hat{\Omega}_{tj}^{-1} \hat{L}_j, \\ \epsilon_{W,n,2} &= \sum_{j=1}^{J_n^d} \mathbf{1}_{n,j} \left\{ \hat{L}_j' \hat{\Omega}_{tj}^{-1} \tilde{\Sigma}_{tj} \hat{\Omega}_{tj}^{-1} (\hat{L}_j - \tilde{L}_j) + (\hat{L}_j - \tilde{L}_j)' \hat{\Omega}_{tj}^{-1} \tilde{\Sigma}_{tj} \hat{\Omega}_{tj}^{-1} \tilde{L}_j \right\}, \\ \epsilon_{W,n,3} &= \sum_{j=1}^{J_n^d} \mathbf{1}_{n,j} \tilde{L}_j' \hat{\Omega}_{tj}^{-1} \tilde{\Sigma}_{tj} \hat{\Omega}_{tj}^{-1} \left(\frac{1}{nq_j} \sum_{i=1}^n R_j(X_i)D_{t,i} \left(\frac{1}{e_t(X_i)} - R_j(X_i)' \gamma_{e,j} \right) \right), \\ \epsilon_{W,n,4} &= \sum_{j=1}^{J_n^d} \mathbf{1}_{n,j} \left(\frac{1}{nq_j} \sum_{i=1}^n R_j(X_i)' D_{t,i} \left(\frac{1}{e_t(X_i)} - R_j(X_i)' \gamma_{e,j} \right) \right)' \hat{\Omega}_{tj}^{-1} \tilde{\Sigma}_{tj} \gamma_{e,j}, \\ \epsilon_{W,n,5} &= \frac{1}{n} \sum_{j=1}^{J_n^d} \mathbf{1}_{n,j} \sum_{i=1}^n \mathbf{1}_{P_j}(X_i) D_{t,i} \varepsilon_{t,i}^2 \left(R_j(X_i)' \gamma_{e,j} - \frac{1}{e_t(X_i)} \right) \left(R_j(X_i)' \gamma_{e,j} + \frac{1}{e_t(X_i)} \right). \end{aligned}$$

Now, $|\epsilon_{W,n,1}| \leq |\epsilon_{W,n,11}| + |\epsilon_{W,n,12}| = o_p(1)$, where applying (A-7),

$$\begin{aligned} |\epsilon_{W,n,11}| &= \left| \sum_{j=1}^{J_n^d} \mathbf{1}_{n,j} \hat{L}_j' \hat{\Omega}_{tj}^{-1} \left(\frac{1}{n} \sum_{i=1}^n 2R_j(X_i)R_j(X_i)' D_{t,i} \varepsilon_i (\hat{\mu}_{tj}(X_i) - \mu_t(X_i)) \right) \hat{\Omega}_{tj}^{-1} \hat{L}_j \right| \\ &\leq C \max_{1 \leq j \leq J_n^d} |\hat{L}_j|^2 \max_{1 \leq j \leq J_n^d} |\mathbf{1}_{n,j} \hat{\Omega}_{tj}^{-1}|^2 \sup_{x \in \mathcal{X}} |\hat{\mu}_t(x) - \mu(x)| \sum_{j=1}^{J_n^d} \sum_{i=1}^n \mathbf{1}_{P_j}(X_i) D_{t,i} |\varepsilon_{t,i}| / n = o_p(1), \end{aligned}$$

and similarly $|\epsilon_{W,n,12}| = o_p(1)$.

Next, $\epsilon_{W,n,2} = o_p(1)$ because

$$\begin{aligned} |\epsilon_{W,n,2}| &\leq 2 \left(\max_{1 \leq j \leq J_n^d} |\hat{L}_j| + |\tilde{L}_j| \right) \left(\max_{1 \leq j \leq J_n^d} |\mathbf{1}_{n,j} \hat{\Omega}_{tj}^{-1}|^2 \right) \left(\max_{1 \leq j \leq J_n^d} \left| \frac{1}{nq_j} \sum_{i=1}^n R_j(X_i) \left(1 - \frac{D_{t,i}}{e_t(X_i)} \right) \right| \right) \\ &\quad \times \sum_{j=1}^{J_n^d} \sum_{i=1}^n |R_j(X_i) R_j(X_i)' D_{t,i} \varepsilon_i^2| / n \\ &= O_p(1) O_p(1) O_p \left(\left(\frac{J_n^d \log J_n}{n} \right)^{1/2} \right) O_p(1) = o_p(1), \end{aligned}$$

where

$$\mathbb{E} \left[\sum_{j=1}^{J_n^d} \sum_{i=1}^n |R_j(X_i) R_j(X_i)' D_{t,i} \varepsilon_i^2| / n \right] \leq C \left(\sup_{x \in \mathcal{X}} |R(x)|^2 \right) \frac{1}{n} \sum_{j=1}^{J_n^d} \sum_{i=1}^n \mathbb{E} [\mathbf{1}_{P_j}(X_i)] = O(1).$$

Next, $\epsilon_{W,n,3} = o_p(1)$ because

$$\begin{aligned} |\epsilon_{W,n,3}| &\leq \left(\max_{1 \leq j \leq J_n^d} |\tilde{L}_j| \right) \left(\max_{1 \leq j \leq J_n^d} |\mathbf{1}_{n,j} \hat{\Omega}_{tj}^{-1}|^2 \right) \\ &\quad \times \sum_{j=1}^{J_n^d} \frac{1}{n^2 q_j} \sum_{i=1}^n \sum_{l=1}^n |R_j(X_l) R_j(X_l)' | D_{t,l} \varepsilon_l^2 \left| R_j(X_i) \left(\frac{1}{e_t(X_i)} - R_j(X_i)' \gamma_{e,j} \right) \right| \\ &= O_p \left(J_n^{-K \wedge S_e} \right) = o_p(1), \end{aligned}$$

since

$$\begin{aligned} &\mathbb{E} \left[\sum_{j=1}^{J_n^d} \frac{1}{n^2 q_j} \sum_{i=1}^n \sum_{l=1}^n |R_j(X_l) R_j(X_l)' | D_{t,l} \varepsilon_l^2 \left| R_j(X_i) \left(\frac{1}{e_t(X_i)} - R_j(X_i)' \gamma_{e,j} \right) \right| \right] \\ &= O_p \left(\left(1 + \frac{J_n^d}{n} \right) J_n^{-K \wedge S_e} \right) = O \left(J_n^{-K \wedge S_e} \right). \end{aligned}$$

Identical reasoning shows $|\epsilon_{W,n,4}| = o_p(1)$ and $|\epsilon_{W,n,5}| = o_p(1)$. Hence $\hat{V}_{W,[t,s]} = V_{W,[t,s]} + o_p(1)$, as $\mathbb{P}(\min_{1 \leq j \leq J_n^d} \mathbf{1}_{n,j} = 1) \rightarrow 1$.

Now consider the “between” term of the variance estimator. For $\hat{V}_{B,[t,s]}$, note that

$$\begin{aligned}\hat{V}_{B,[t,s]} &= \frac{1}{n} \sum_{i=1}^n \hat{\mu}_t(X_i) \hat{\mu}_s(X_i) - \hat{\mu}_s \frac{1}{n} \sum_{i=1}^n \hat{\mu}_t(X_i) - \hat{\mu}_t \frac{1}{n} \sum_{i=1}^n \hat{\mu}_s(X_i) + \hat{\mu}_s \hat{\mu}_t \\ &= \frac{1}{n} \sum_{i=1}^n \mu_t(X_i) \mu_s(X_i) - \hat{\mu}_s \frac{1}{n} \sum_{i=1}^n \mu_t(X_i) - \hat{\mu}_t \frac{1}{n} \sum_{i=1}^n \mu_s(X_i) + \hat{\mu}_s \hat{\mu}_t \\ &\quad + \epsilon_{B,n,1} + \epsilon_{B,n,2} + \epsilon_{B,n,3} + \epsilon_{B,n,4} + \epsilon_{B,n,5},\end{aligned}$$

where

$$\begin{aligned}\epsilon_{B,n,1} &= \frac{1}{n} \sum_{i=1}^n (\hat{\mu}_t(X_i) - \mu_t(X_i)) (\hat{\mu}_s(X_i) - \mu_s(X_i)), \\ \epsilon_{B,n,2} &= \frac{1}{n} \sum_{i=1}^n \mu_t(X_i) (\hat{\mu}_s(X_i) - \mu_s(X_i)), \quad \epsilon_{B,n,3} = \frac{1}{n} \sum_{i=1}^n (\hat{\mu}_t(X_i) - \mu_t(X_i)) \mu_s(X_i), \\ \epsilon_{B,n,4} &= -\hat{\mu}_s \frac{1}{n} \sum_{i=1}^n (\hat{\mu}_t(X_i) - \mu_t(X_i)), \quad \epsilon_{B,n,5} = -\hat{\mu}_t \frac{1}{n} \sum_{i=1}^n (\hat{\mu}_s(X_i) - \mu_s(X_i)).\end{aligned}$$

Thus, because $\hat{\mu} - \mu = o_p(1)$ and Result (A-7) holds under the assumptions of the theorem, $\epsilon_{B,n,k} = o_p(1)$ for $k = 1, \dots, 5$, and $\hat{V}_{B,[t,s]} = V_{B,[t,s]} + o_p(1)$ as stated. \blacksquare

REFERENCES

- ABADIE, A., D. DRUKKER, J. L. HERR, AND G. W. IMBENS (2004): “Implementing Matching Estimators for Average Treatment Effects in Stata,” *The Stata Journal*, 4(3), 290–311.
- ABADIE, A., AND G. W. IMBENS (2006): “Large Sample Properties of Matching Estimators for Average Treatment Effects,” *Econometrica*, 74(1), 235–267.
- BANG, H., AND J. M. ROBINS (2005): “Doubly Robust Estimation in Missing Data and Causal Inference Models,” *Biometrics*, 61, 962–972.
- CATTANEO, M. D. (2010): “Efficient Semiparametric Estimation of Multi-valued Treatment Effects under Ignorability,” *Journal of Econometrics*, 155(2), 138–154.
- CATTANEO, M. D., AND M. FARRELL (2011): “Optimal Convergence Rates, Bahadur Representation, and Asymptotic Normality of Partitioning Estimators,” working paper.
- CATTANEO, M. D., G. W. IMBENS, C. PINTO, AND G. RIDDER (2009): “Subclassification on the Propensity Score: Large Sample Properties,” work in progress.
- CHEN, X., H. HONG, AND A. TAROZZI (2004): “Semiparametric Efficiency in GMM Models of Nonclassical Measurement Errors, Missing Data and Treatment Effects,” Cowles Foundation Discussion Paper No. 1644.
- (2008): “Semiparametric Efficiency in GMM Models With Auxiliary Data,” *Annals of Statistics*, 36(2), 808–843.
- COCHRAN, W. G. (1968): “The Effectiveness of Adjustment by Subclassification in Removing Bias in Observational Studies,” *Biometrics*, 24(2), 295–313.
- FIRPO, S. (2007): “Efficient Semiparametric Estimation of Quantile Treatment Effects,” *Econometrica*, 75(1), 259–276.
- GYÖRFI, L., M. KOHLER, A. KRZYŻAK, AND H. WALK (2002): *A Distribution-Free Theory of Nonparametric Regression*. Springer, New York.
- HAHN, J. (1998): “On the Role of the Propensity Score in Efficient Semiparametric Estimation of Average Treatment Effects,” *Econometrica*, 66(2), 315–331.
- HECKMAN, J. J., AND E. J. VYTLACIL (2007): “Econometric Evaluation of Social Programs, Part I: Causal Models, Structural Models and Econometric Policy Evaluation,” in *Handbook of Econometrics*, vol. VI, ed. by J. Heckman, and E. Leamer, pp. 4780–4874. Elsevier Science B.V.
- HIRANO, K., G. W. IMBENS, AND G. RIDDER (2003): “Efficient Estimation of Average Treatment Effects using the Estimated Propensity Score,” *Econometrica*, 71(4), 1161–1189.
- HOLLAND, P. W. (1986): “Statistics and Causal Inference,” *Journal of the American Statistical Association*, 81(396), 945–960.

- IMAI, K., AND D. A. VAN DYK (2004): “Causal Inference With General Treatment Regimes: Generalizing the Propensity Score,” *Journal of the American Statistical Association*, 99(467), 854–866.
- IMBENS, G. W. (2000): “The Role of the Propensity Score in Estimating Dose-Response Functions,” *Biometrika*, 87(3), 706–710.
- (2004): “Nonparametric Estimation of Average Treatment Effects under Exogeneity: A Review,” *Review of Economics and Statistics*, 86(1), 4–29.
- IMBENS, G. W., W. K. NEWEY, AND G. RIDDER (2006): “Mean-Squared-Error Calculations for Average Treatment Effects,” Working Paper.
- IMBENS, G. W., AND J. M. WOOLDRIDGE (2009): “Recent Developments in the Econometrics of Program Evaluation,” *Journal of Economic Literature*, 47(1), 5–86.
- LECHNER, M. (2001): “Identification and estimation of causal effects of multiple treatments under the conditional independence assumption,” in *Econometric Evaluations of Active Labor Market Policies*, ed. by M. Lechner, and E. Pfeiffer, pp. 43–58. Physica, Heidelberg.
- ROSENBAUM, P. R., AND D. B. RUBIN (1983): “The Central Role of the Propensity Score in Observational Studies for Causal Effects,” *Biometrika*, 70(1), 41–55.
- (1984): “Reducing Bias in Observational Studies Using Subclassification on the Propensity Score,” *Journal of the American Statistical Association*, 79(1), 516–524.
- TSIATIS, A. A. (2006): *Semiparametric Theory and Missing Data*. Springer, New York.
- WOOLDRIDGE, J. M. (2007): “Inverse probability weighted estimation for general missing data problems,” *Journal of Econometrics*, 141, 1281–1301.

Figure 1: Empirical Average Bias for Univariate Models

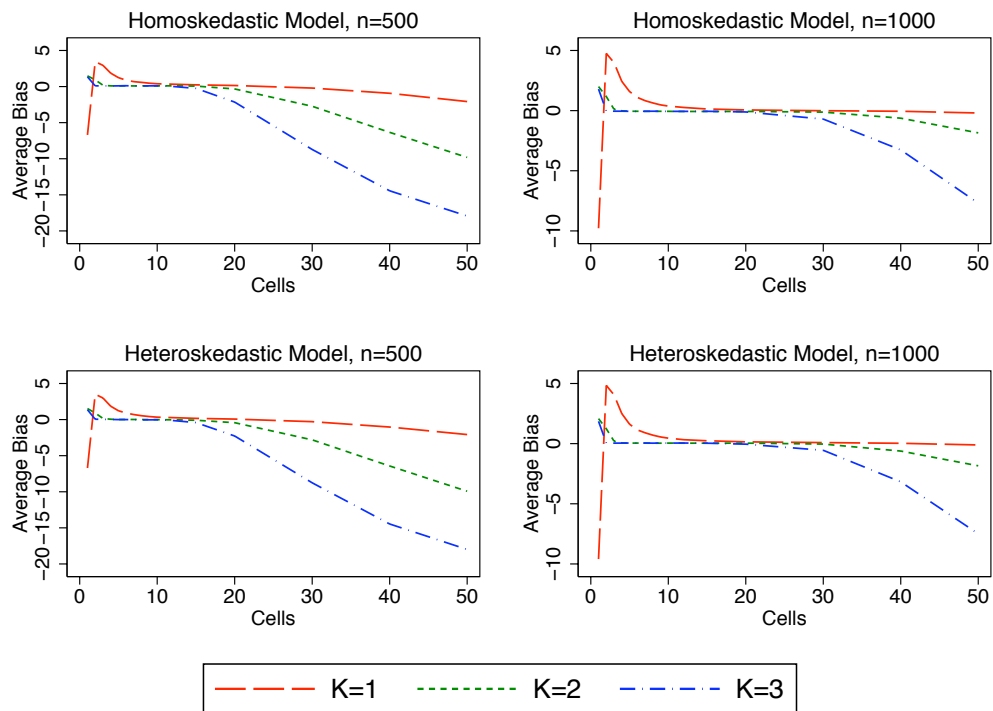


Figure 2: Coverage Rates for 95% Confidence Intervals for Univariate Models

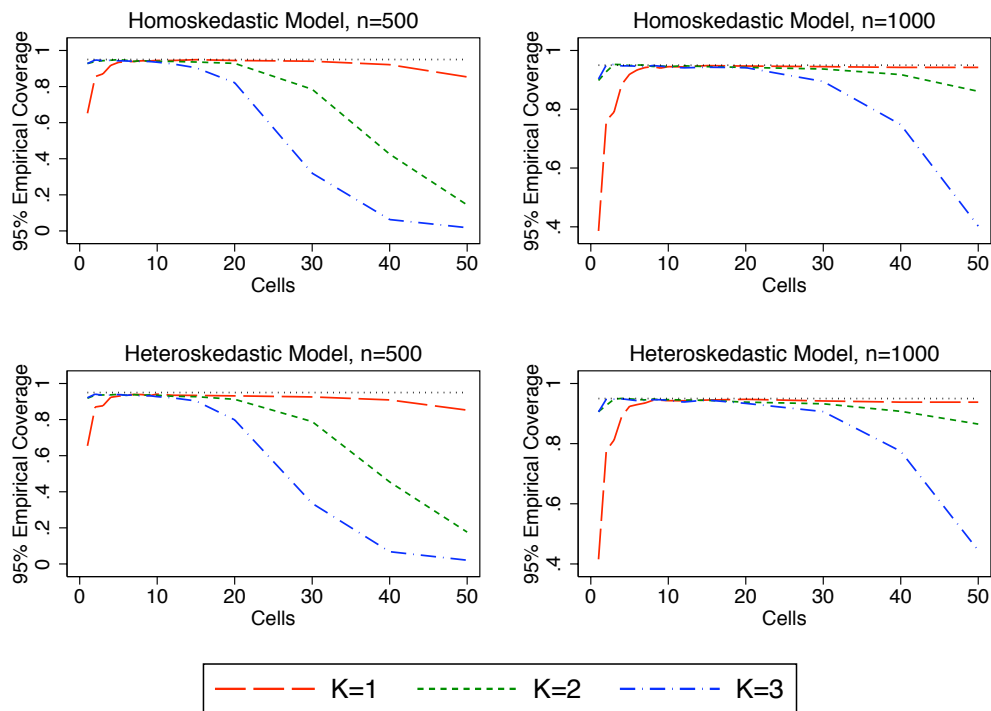


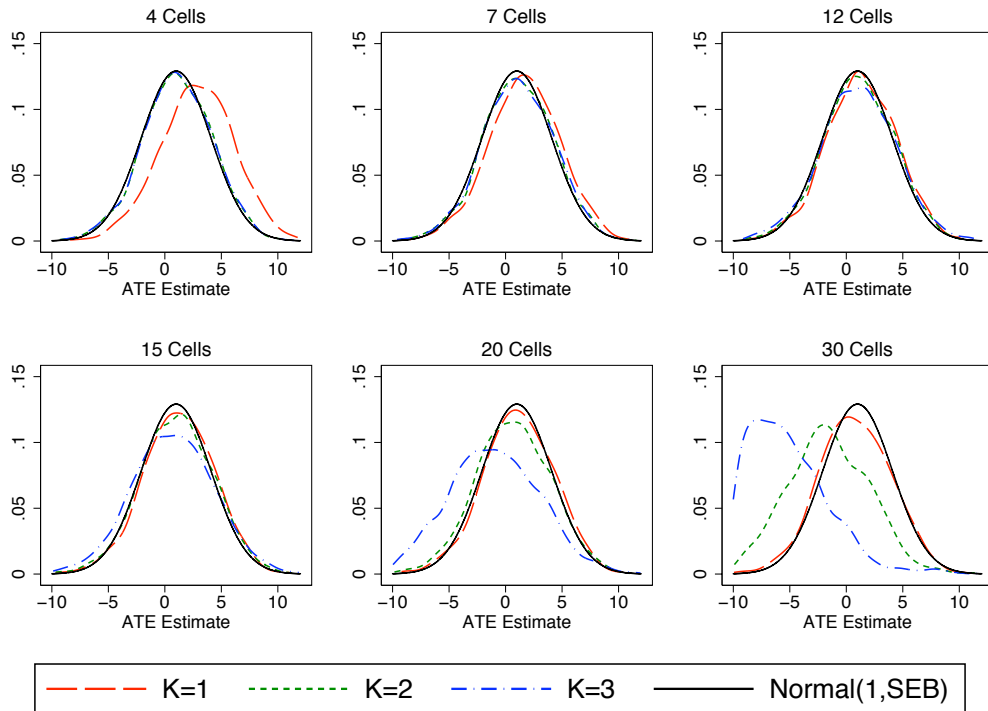
Figure 3: Gaussian Approximation for Univariate Homoskedastic Model, $n = 500$ 

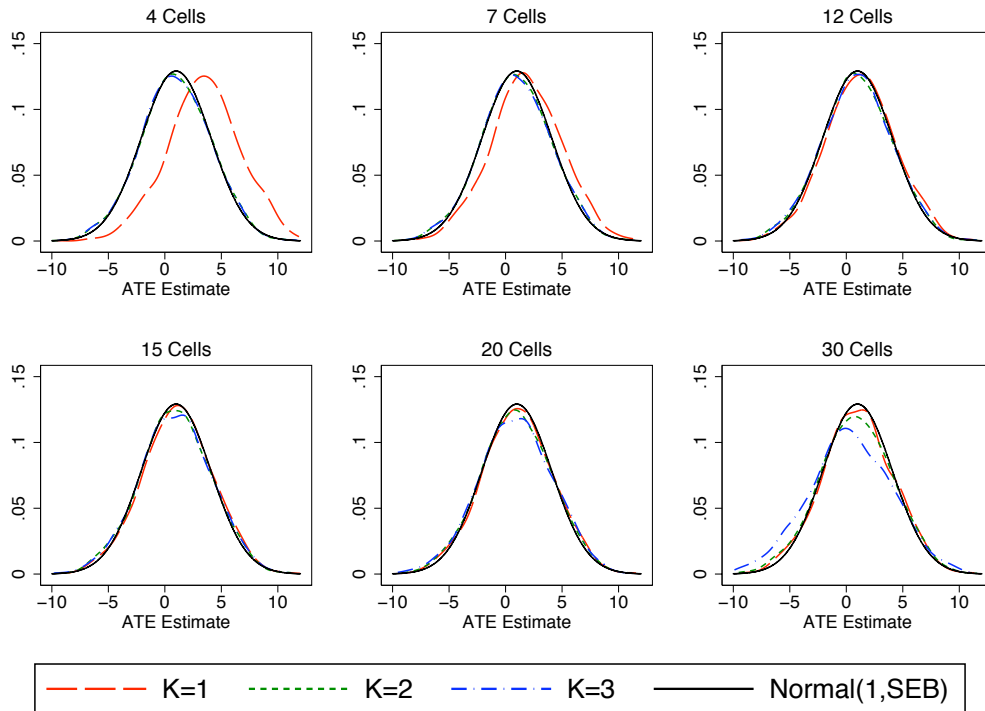
Figure 4: Gaussian Approximation for Univariate Homoskedastic Model, $n = 1,000$ 

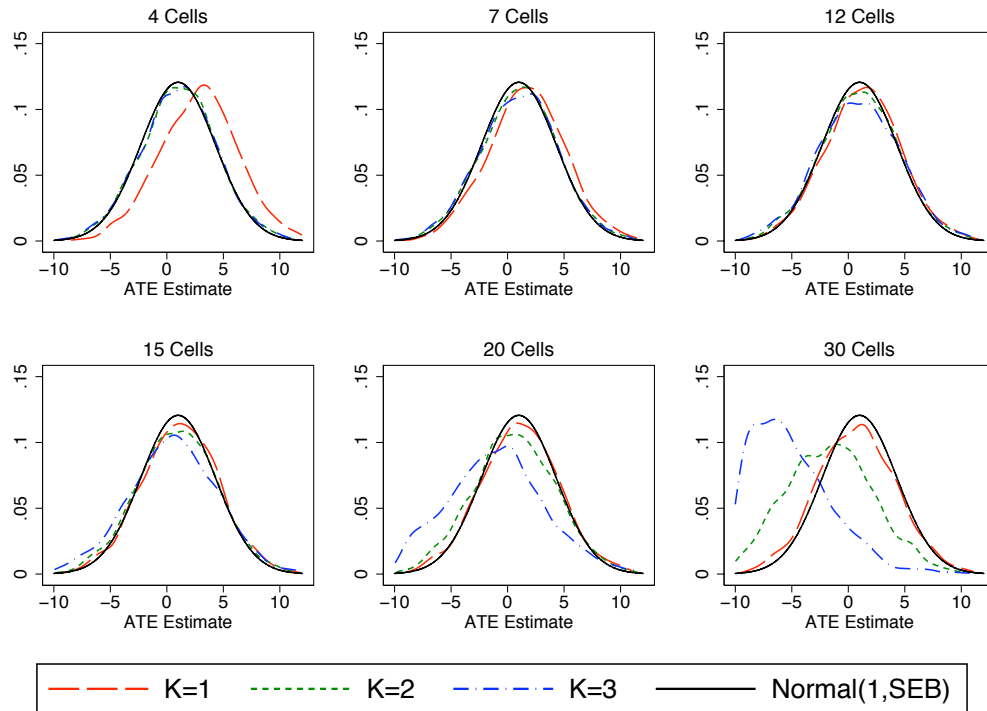
Figure 5: Gaussian Approximation for Univariate Heteroskedastic Model, $n = 500$ 

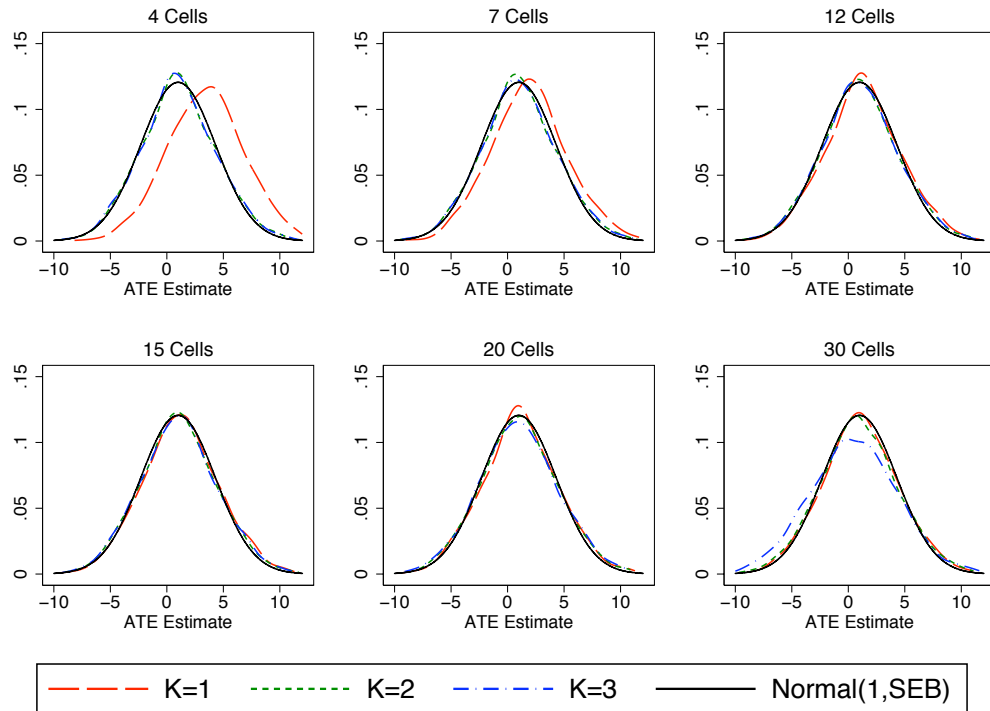
Figure 6: Gaussian Approximation for Univariate Heteroskedastic Model, $n = 1,000$ 

Figure 7: Empirical Average Bias for Bivariate Models

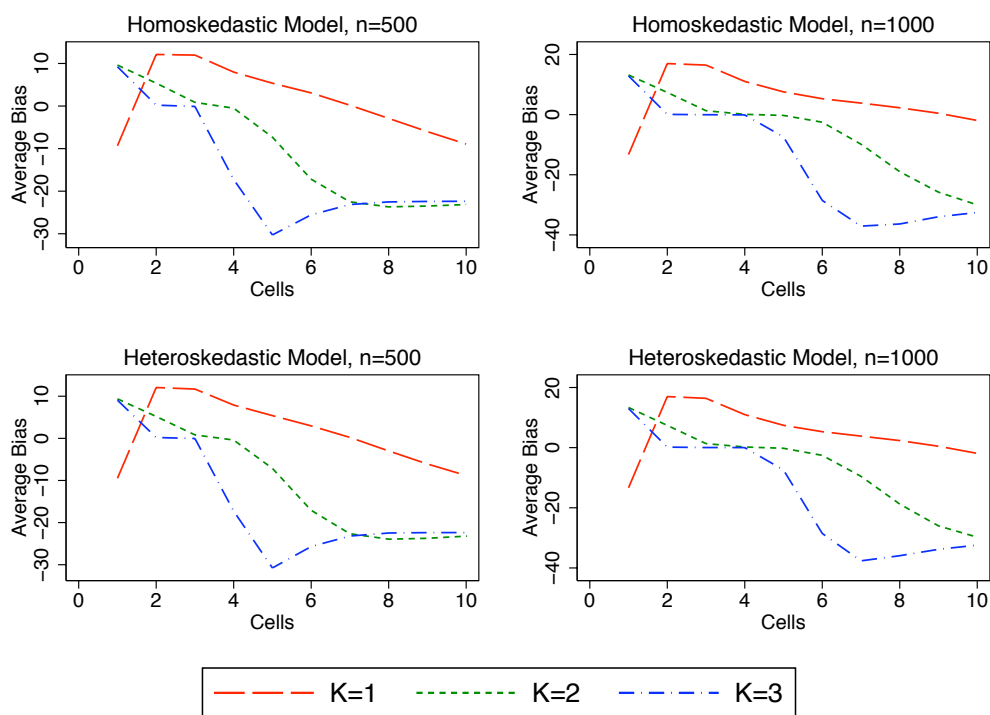


Figure 8: Coverage Rates for 95% Confidence Intervals for Bivariate Models

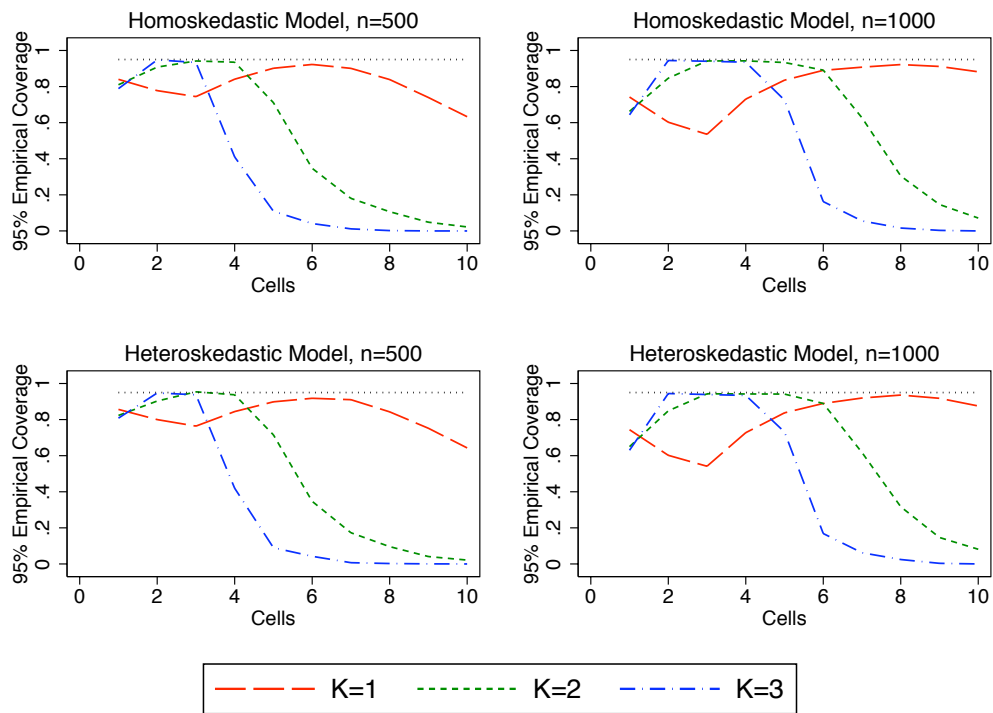


Figure 9: Gaussian Approximation for Bivariate Homoskedastic Model, $n = 500$

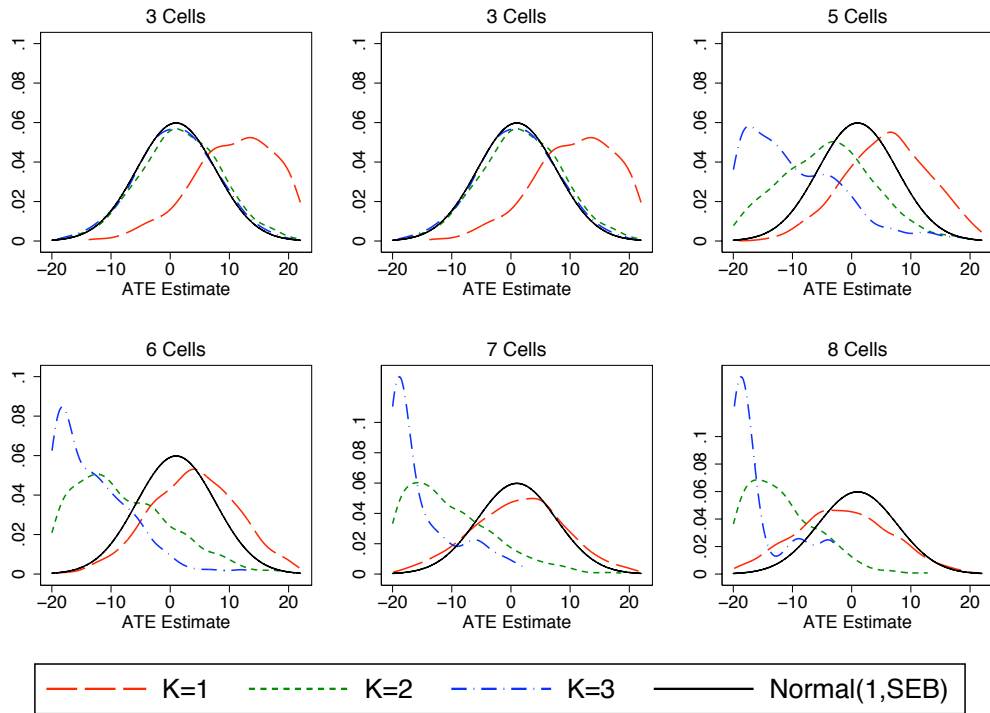


Figure 10: Gaussian Approximation for Bivariate Homoskedastic Model, $n = 1,000$

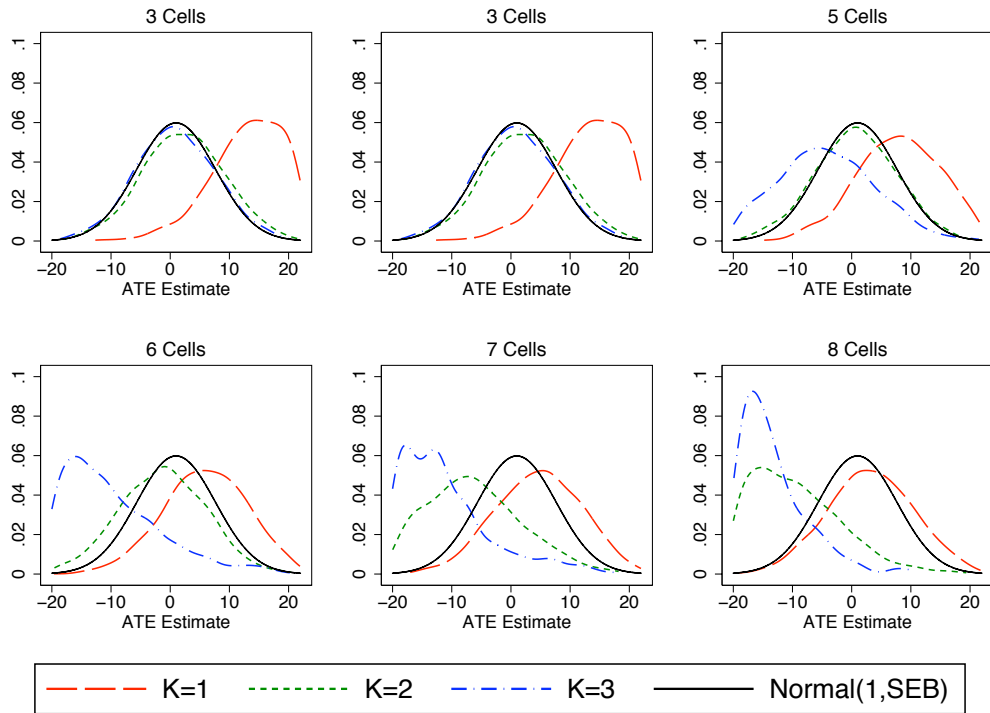


Figure 11: Gaussian Approximation for Bivariate Heteroskedastic Model, $n = 500$

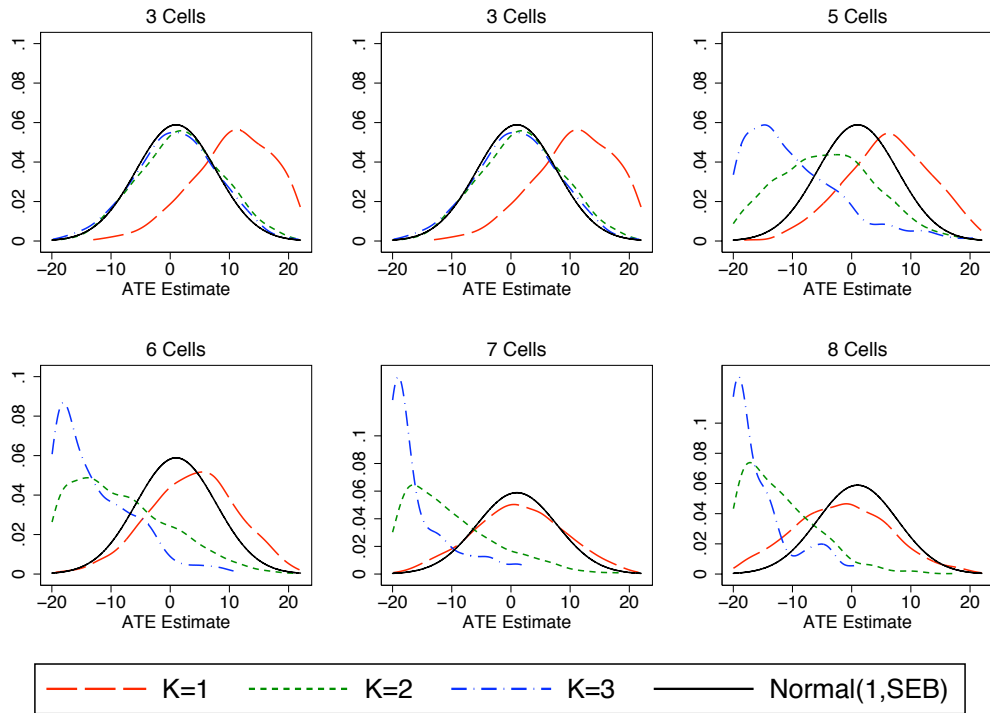


Figure 12: Gaussian Approximation for Bivariate Heteroskedastic Model, $n = 1,000$

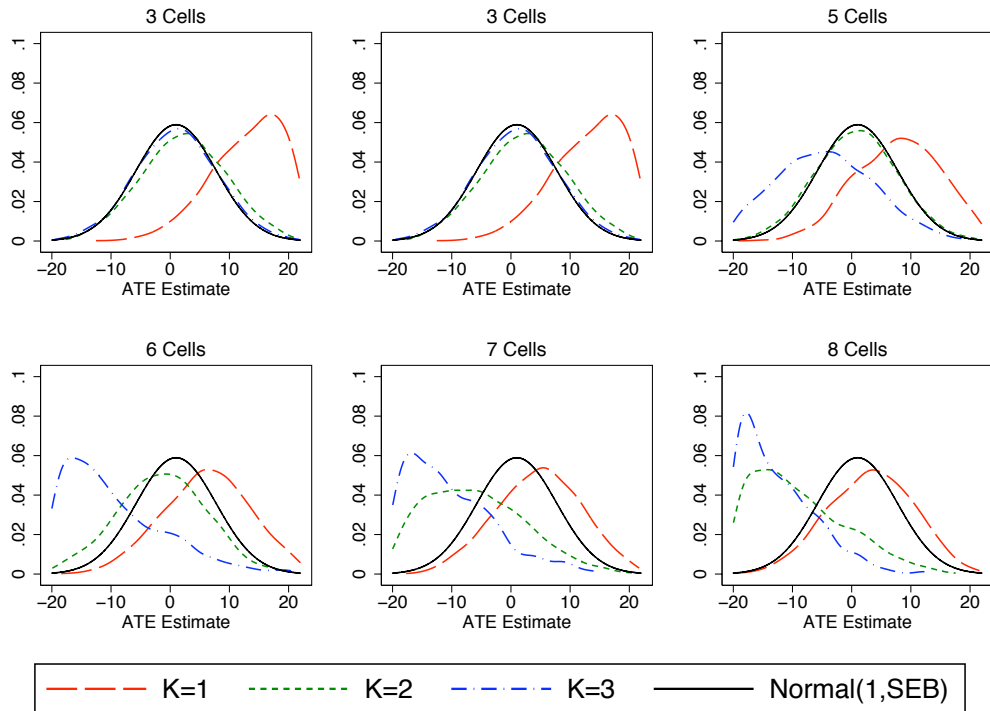


Figure 13: Gaussian Approximation for Bivariate Homoskedastic Model, $n = 2,000$

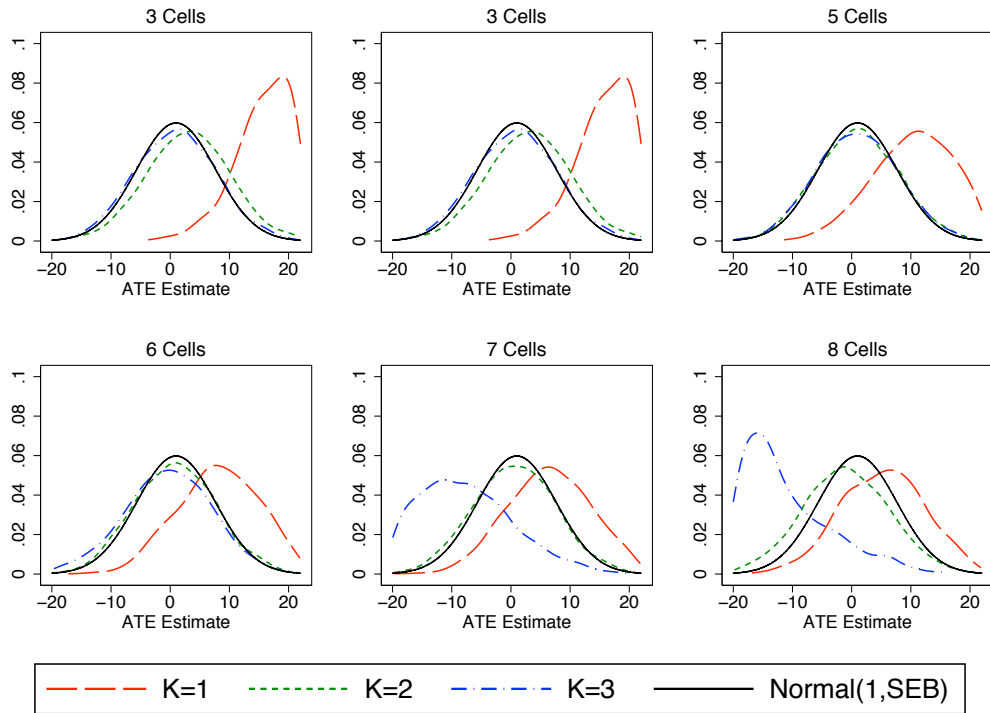


Table 1: Mean-Square Error Compared to Alternative Estimators

		$n = 500$			$n = 1,000$		
		Bias	Var.	MSE	Bias	Var.	MSE
Subclassification Estimator							
$J_n = 7$	$K = 1$	0.484	10.172	10.407	0.887	9.868	10.655
$J_n = 7$	$K = 2$	-0.14	10.087	10.107	0.015	9.474	9.474
$J_n = 10$	$K = 1$	0.174	10.11	10.14	0.427	9.543	9.726
$J_n = 10$	$K = 2$	-0.138	10.285	10.304	0.021	9.595	9.595
IPW							
	Degree 4	-0.137	9.902	9.921	0.016	9.405	9.405
	Degree 6	-0.358	10.567	10.695	-0.16	9.738	9.764
Series							
	Degree 4	-0.136	9.804	9.822	0.012	9.35	9.35
	Degree 6	-0.357	10.446	10.573	-0.169	9.725	9.754
NN-Matching							
Simple	M=1	-0.135	13.24	13.258	-0.003	12.619	12.619
Simple	M=2	-0.113	11.501	11.514	0.021	10.982	10.982
Bias-adj	M=1	-0.138	13.24	13.259	-0.005	12.618	12.618
Bias-adj	M=2	-0.12	11.502	11.516	0.018	10.981	10.981

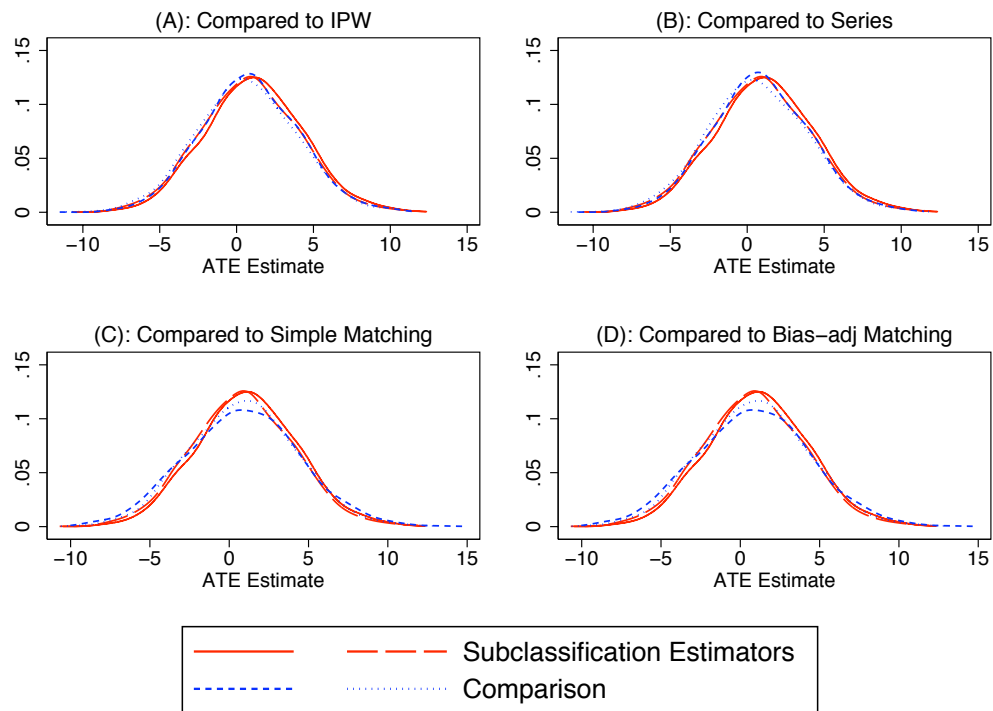
Figure 14: Comparison to Alternative Estimators, $n = 500$ 

Figure 15: Comparison to Alternative Estimators, $n = 1,000$ 