Kitagawa, Toru; Muris, Chris

Working Paper

Covariate selection and model averaging in semiparametric estimation of treatment effects

cemmap working paper, No. CWP61/13

Provided in Cooperation with:
Institute for Fiscal Studies (IFS), London

Suggested Citation: Kitagawa, Toru; Muris, Chris (2013) : Covariate selection and model averaging in semiparametric estimation of treatment effects, cemmap working paper, No. CWP61/13, Centre for Microdata Methods and Practice (cemmap), London, https://doi.org/10.1920/wp.cem.2013.6113

This Version is available at:
https://hdl.handle.net/10419/97401

cemmap
centre for microdata methods and practice

Covariate selection and model averaging in semiparametric estimation of treatment effects

Toru Kitagawa
Chris Muris

The Institute for Fiscal Studies
Department of Economics, UCL

cemmap working paper CWP61/13

E·S·R·C ECONOMIC & SOCIAL RESEARCH COUNCIL
An ESRC Research Centre

Covariate Selection and Model Averaging in Semiparametric Estimation of Treatment Effects*

Toru Kitagawa† and Chris Muris‡

December 2, 2013

Abstract

In the practice of program evaluation, choosing the covariates and the functional form of the propensity score is an important choice for estimating treatment effects. This paper proposes data-driven model selection and model averaging procedures that address this issue for the propensity score weighting estimation of the average treatment effects for treated (ATT). Building on the focussed information criterion (FIC), the proposed selection and averaging procedures aim to minimize the estimated mean squared error (MSE) of the ATT estimator in a local asymptotic framework. We formulate model averaging as a statistical decision problem in a limit experiment, and derive an averaging scheme that is Bayes optimal with respect to a given prior for the localisation parameters in the local asymptotic framework. In our Monte Carlo studies, the averaging estimator outperforms the post-covariate-selection estimator in terms of MSE, and shows a substantial reduction in MSE compared to conventional ATT estimators. We apply the procedures to evaluate the effect of the labor market program described in LaLonde (1986).

Keywords: Treatment effects, Propensity score, Model selection, Focussed information criterion, Model averaging.

*We thank Debopam Bhattacharya, Irene Botosaru, Xiaohong Chen, Christian Hansen, Yuichi Kitamura, Frank Kleibergen, Michael Lechner, Simon Lee, Richard Smith, and Frank Windmeijer for valuable comments and discussions. We also thank the seminar participants at AMES 2013, University of Bristol, University of British Columbia, Brown University, the Cemmap/PEPA workshop on Program Evaluation, University of Groningen, University of Sankt Gallen, Tilburg University, Toulouse School of Economics, and Yale Econometrics Lunch for their helpful comments. All remaining errors are ours. Financial support from the ESRC through the ESRC Centre for Microdata Methods and Practice (CeMMAP) (grant number RES-589-28-0001) is gratefully acknowledged.

†Department of Economics, University College London. Email: t.kitagawa@ucl.ac.uk.

‡Department of Economics, Simon Fraser University. Email: cmuris@sfu.ca.

1 Introduction

A large body of empirical research in economics is concerned with estimation of the causal impact of various social programs. When the exposure or participation to the policy program is not randomized, researchers often use observational data in conjunction with the assumption that treatment assignment is random once a set of observable pre-treatment covariates is conditioned on (unconfoundedness). Several semi-parametric procedures that rely on the unconfoundedness assumption have been proposed, including propensity score matching (Rosenbaum and Rubin (1983) and Heckman, Ichimura, and Todd (1998)); covariate matching (Abadie and Imbens (2006)); regression (Imbens, Newey, and Ridder (2005)); propensity score weighting (Hirano, Imbens, and Ridder (2003)); and a combination of the latter two (Hahn (1998)). Imbens (2004) provides an excellent review on these methods.

A common concern that arises when using such estimators is that the researcher has to choose which covariates to include as confounders, and which functional form specification is used in modelling the propensity score or/and the outcome equations. The literature on semiparametric estimation has been rather silent on a formal treatment of this practical issue. As a result, empirical researchers using these methods rarely provide formal justification for the chosen specification in reporting the estimation results. Common practice is to conduct some informal sensitivity check by seeing how the estimate changes over different specifications.

The main goal of this paper is to offer a data-driven procedure that gives a best causal effect estimator for the *average treatment effects for treated* (ATT) in the presence of specification uncertainty on the propensity score. We focus on propensity score weighting estimators. As a way to handle specification uncertainty, this paper considers both model selection and model averaging. Building on the idea of focussed information criterion (hereafter FIC) proposed by Claeskens and Hjort (2003), our model selection procedure aims to select the specification of the propensity score that minimizes the mean squared error (hereafter MSE) of the ATT estimator. Along the idea of frequentist model averaging (Hjort and Claeskens (2003)), the model averaging of this paper generalizes the FIC-based model selection. That is, instead of selecting one model and estimating ATT based on the selected model, we consider constructing a point estimator for ATT in the form of a weighted average of the point estimators over the candidate models, where the weights are optimally chosen so as to minimize MSE for the ATT parameter.

The FIC based specification search procedure proposed in this paper works as follows. As an

input of the procedure, the researcher provides a most complicated specification (largest model) of the propensity score in the following parametric form,

$$\Pr(D=1|X) = G\left(W(X)'\gamma\right),$$

where $D = 1$ (treated) or $D = 0$ (control) is an indicator of the treatment status; X is the set of all conditioning covariates considered by the researcher; $W(X)$ is a vector of functions of the regressors X that can contain interactions and nonlinear transformations of X; and $G(\cdot)$ is a known link function such as the logit function. Candidates for the optimal specification are given as submodels of the most complicated specification, where each submodel corresponds to a subset vector of $W(X)$ to be included in propensity score estimation. We assume that the unconfoundedness assumption holds for the full set of covariates X, and that the ATT parameter is identified and consistently estimated by a $\sqrt{n}$-asymptotically normal estimator in this largest model. We assume that the candidate specifications are locally misspecified in the sense that the true values of coefficients γ are in a $n^{-1/2}$-neighborhood of zero with a radius governed by a localization parameter δ. In this local asymptotic framework, we obtain the asymptotic MSE of the ATT estimator in each candidate model. Then, the FIC is then defined as an estimate of the asymptotic MSE of the ATT estimator in each candidate specification, where uncertainty on the localization parameters are treated by either a plug-in manner or Bayesian manner. The procedure yields a model (subvector of $W(X)$) that minimizes the FIC, and the ATT estimator computed in the chosen specification can be reported as a post-selection point estimator for ATT.

As an estimator for the ATT in each candidate model, we employ the normalized propensity score weight (hereafter NPW) estimator (Imbens (2004)). The NPW estimator for ATT has several attractive features compared with the naive propensity score weighted estimator (as in Wooldridge (2002), equation 18.22). The NPW estimator has a smaller asymptotic variance than the simple ATT estimator when a parametric specification for the propensity score is employed. The NPW estimator is simple to implement and simulation evidence suggests excellent finite sample performance of the NPW estimator (Busso, DiNardo, and McCrary (2011)). The main reason that we focus on the ATT rather than ATE closely relates to the fact that the semiparametric efficiency bound for ATT can be improved if knowledge on a specification of the propensity score is available, see Hahn (2004); Chen, Hong, and Tarozzi (2008); and Graham, de Xavier Pinto, and Egel (2011). Using the local asymptotic approximation, the NPW estimator for the ATT in the parsimonious specification can have a smaller asymptotic variance than in the largest model, due to the gain in the efficiency bound for ATT by knowledge of parsimonious specification for the propensity score.

The parsimonious model, on the other hand, can be biased due to the local misspecification. As a result, there is a bias-variance tradeoff, which the FIC-based selection procedure aims to optimally balance out.

The second goal of this paper is to develop a model averaging estimator for ATT in the presence of model uncertainty regarding propensity score specifications. Under the local misspecification framework described above, we consider choosing the model weights so as to minimize the asymptotic MSE of the averaged ATT estimator. The asymptotic MSE to be minimized, however, depends on the unknown localization parameters that cannot be consistently estimated. In order to deal with the non-vanishing uncertainty for the localization parameters, we pose the problem of choosing optimal weights as a statistical decision problem in the limit Gaussian experiment (see e.g. Chapter 7 of van der Vaart (1998)). We then derive the optimal weights in the sense of Bayes decision in the limit experiment with respect to a prior for the localization parameters. Our approach to the optimal averaging weights leads to a weighting scheme that is different from the plug-in based procedure and the inverse-FIC based weights of Hjort and Claeskens (2003), whose treatment of the localization parameters, to the best of our knowledge, lacks a decision-theoretic optimality argument.

We conduct Monte Carlo studies in order to examine the finite sample performance of the proposed procedures. Our Monte Carlo results show that the model averaging estimator outperforms in terms of MSE the FIC-based post-selection NPW estimator and the NPW estimators in any of the candidate models including the MSE minimizing one. The MSE gain of implementing the model averaging estimator can be substantial relative to a correctly specified large model; in our Monte Carlo specification, the model averaging estimator improves the MSE of a correctly specified largest model by 20-30%. To illustrate the use of our model selection and averaging procedures, we apply them to the Lalonde's (1986) data of a job-training program in the US.

1.1 Related Literatures

In contrast to the likelihood based model selection procedures such as AIC or BIC, the focussed information criterion (FIC) aims to select a best model in terms of estimation accuracy for a focussed parameter. Importance of this focussed view in model selection is also emphasized in Hansen (2005) in the time series context. Several authors have extended the FIC idea to the model selection problems in various semiparametric models; for instance, Hjort and Claeskens

(2006), Claeskens and Carroll (2007), and Zhang and Liang (2011), among others. We believe that the focussed viewpoint is particularly appealing for covariate selection in the program evaluation context since many program evaluation studies have a well-defined parameter of interest. We share this view with Vansteelandt, Bekaert, and Claeskens (2012), in which they propose a FIC-based variable selection procedure for a parametric binary regression model with log-odds ratio as a focussed parameter. This paper, in contrast, analyzes the variable selection problem for propensity score estimation with a focussed parameter being the average treatment effects for treated, and our approach does not require parametric specifications for the outcome regression equations.

The problem of specification choice in the semiparametric estimation procedure, using nonparametric estimators of the propensity scores or/and the outcome regression equations, is reduced to the problem of how to select smoothing parameters, such as the kernel bandwidth or the number of terms in series regression. To our knowledge, Ichimura and Linton (2001) and Imbens et al. (2005) are the only works that discuss the choice of smoothing parameters with focusing on minimizing the MSE of the semiparametric ATE estimator. Compared with their approach, our approach is "less non-parametric", in the sense that our approach imposes a parametric restriction on the propensity score in the largest model. Positive consequences of the parametric specification are that we can deal with multi-dimensional covariates in a simple manner, and that the proposed procedure does not require a preliminary nonparametric estimate of unknown functions (cf. Ichimura and Linton (2001)). On the other hand, our approach relies on a user-specified largest model, and is not free from the arbitrariness concern in the choice of largest model. A similar concern would also arise in the procedure of Imbens et al. (2005), in which a choice of basis functions as well as their ordering are important inputs specified by the user.

The l_1-penalized likelihood procedure (Lasso) proposed by Tibshirani (1996) is a powerful tool in the variable selection context, especially when the number of candidate regressors is large. Belloni, Chernozhukov, and Hansen (2013) recently develop the so-called double-selection lasso method for covariate selection and post-selection inference for estimation of various treatment effects in the presence of high-dimensional covariates. Our FIC approach to covariate selection differs from their Lasso approach in terms of the scope for applications and the notion of optimality that these procedures aim to achieve asymptotically. First, our FIC procedure mainly targets at the situations, where the number of regressors is much smaller than the sample size, while, with employing the sparsity restrictions, the Lasso approach can effectively handle the situations, where the number of regressors is as many as or even larger than the sample size. Second, optimality of the FIC-

based covariate selection and averaging hinges on a decision theoretic optimality in a limit Gaussian experiment, while theoretical justification of the Lasso-based covariate selection approach invokes the oracle property. Third, this paper mainly focuses on model selection and averaging for point estimation, and provides little contribution to post-selection inference, whereas, as one of their remarkable contributions, Belloni et al. (2013) demonstrate that post-selection inference with their Lasso procedures yields a uniformly valid inference procedure for ATEs and ATTs.

The model averaging considered in this paper stands on the frequentist model averaging viewpoint; the procedure aims to find an averaging weight that optimizes accuracy of the averaged estimator in terms of the MSE criterion of the ATT parameter. The frequentist model averaging targeted at the MSE for a focussed parameter is pursued by Hjort and Claeskens (2003) in general parametric models. This paper extends their model averaging framework to the context of semiparametric estiamtion of ATT. For variable selection problem in the least squares context, frequentist model averaging with the MSE criterion of the entire regression function (integrated MSE) is analyzed by Hansen (2007), Wan, Zhang, and Zou (2010), and Hansen and Racine (2012). Magnus, Powell, and Prüfer (2010) proposes a way of designing a prior in the Bayesian model averaging based on the frequentist considerations of the mean squared errors. See also Hjort and Claeskens (2008) for an overview of model averaging and further references. DiTraglia (2013) and Sueishi (2013) extend the parametric framework of Hjort and Claeskens (2003) to semiparametric models defined by a set of moment conditions, and develop the FIC and FIC-based model averaging for generalized method of moment estimators, with primary applications to linear instrumental variable models. Lu (2013) considers averaging semiparametric estimators for ATE or ATT in a manner similar to the frequentist model averaging of Hjort and Claeskens (2003), where the estimator in each model uses nonparametrically estimated regression or propensity score functions with a different set of conditioning covariates. In contrast to the approach of Lu (2013), our approach concerns not only a choice of covariates, but also a functional form specification of the propensity scores.

Being different from the Hjort and Claeskens's proposal of the plug-in based averaging weights, our derivation of the optimal averaging weights solves a Bayes optimal statistical decision in a limit normal experiment. In econometrics, decision-theoretic analyses in limit experiments are conducted in various contexts; see Hirano and Porter (2009) for the treatment choice problem, and Song (2013) for point estimation problem for intervally-identified models.

1.2 Plan of the Paper

In Section 2, we introduce the local misspecification framework in the ATT estimation, and derive the asymptotic MSE of the candidate models. We also examine the variance-bias trade-offs between large and parsimonious models through the analytical expression of the asymptotic MSE. In Section 3, we derive the FIC and propose the FIC based covariate selection procedure. We also extend the FIC-based model selection analysis to model averaging. Monte Carlo studies are provided in Section 4 to examine the performance of the proposed model selection and model averaging procedures. Section 5 applies our model selection and averaging procedure to the Lalonde's (1986) data set on the National Supported Work Demonstration job-training program. Section 6 concludes. All proofs of the proposition and auxiliary lemmas are collected in Appendix A.

2 Semiparametric Estimation of ATT with Local Misspecification

Let $\{(Y_i, D_i, X_i') : i = 1, \ldots, n\}$ be a size n random sample where an observation consists of a scalar observed outcome $Y_i \in \mathcal{R}$, a binary treatment status $D_i \in \{0, 1\}$, and a (column) vector of covariates $X_i \in \mathbf{X}$. Suppose that we have L predetermined covariates available for every individual in the sample, $X_i' = (X_{i1}, \ldots, X_{iL})$. Each covariate can be either discrete or continuous. We denote the potential outcomes corresponding to each treatment status as $Y_i(1)$ and $Y_i(0)$. The observed outcome Y_i is linked to the potential outcomes through $Y_i = D_i Y_i(1) + (1 - D_i) Y_i(0)$. The estimand of interest focussed in this paper is the population average treatment effects for treated (ATT), $\tau = E(Y(1) - Y(0)|D = 1)$.

The specification selection procedure to be proposed in this paper first asks the researcher to specify a most complicated specification for the propensity score function. We refer to the model with this most complicated specification of the propensity score as the *largest model*. Let $W(X) \in \mathcal{R}^K$ be a vector of regressors with length K that is to be included in the propensity score estimation in the largest model. $W(X)$ includes an intercept, and may contain interactions and nonlinear transformations of X. In the subsequent asymptotic analysis, we will *not* let its dimension K grow with the sample size. In practical terms, the fixed dimension of $W(X)$ means that the number of regressors in the largest model is specified to be relatively small compared to the sample size. We use a short-hand notation, $W_i = W(X_i)$, as far as no confusion arises.

Let $\mathcal{P}$ be the set of distributions of $(Y_i(1), Y_i(0), D_i, X_i)$ that contains the true data generating process. We impose the following set of assumptions concerning the set of population distributions

$\mathcal{P}$.

Assumption 2.1:

(i) *(Unconfoundedness) For every* $P \in \mathcal{P}$*, the potential outcomes* $(Y(1), Y(0))$ *are conditionally independent of treatment status* $D \in \{1, 0\}$ *given the full set of covariates* $X \in \mathbb{R}^L$*.*

(ii) *(Correct specification of propensity score) For every* $P \in \mathcal{P}$*, there exists a unique* $\gamma(P) \in \mathbb{R}^K$ *such that* $P(D = 1|X) = G\left(W(X)'\gamma(P)\right)$ *for a known monotone twice continuously differentiable link function* $G(\cdot)$ *common to all* $P \in \mathcal{P}$*.*

(iii) *(Strict overlap) There is* $\kappa > 0$ such that $\kappa \leq G\left(W(X)'\gamma(P)\right) \leq 1 - \kappa < 1$ holds for all X in the support of X, uniformly over $P \in \mathcal{P}$.

The first part of Assumption 2.1 combined with the overlap condition ensures that the ATT is identifiable in the largest model. Since the unconfoundedness assumption is not testable, the covariate selection procedure of this paper is unable to assess validity of the unconfoundedness assumption. The second part of Assumption 2.1 states that the propensity score has a parametric single index structure with a known link function. The literature of semiparametric estimation of ATT commonly introduces nonparametric propensity scores (e.g., Hahn (1998), Hirano et al. (2003)), while we restrict our analysis to the case with parametric propensity scores. This assumption may appear restrictive at a theoretical level, but does not bind much in empirical practice, since, with a finite number of observations, implementation of nonparametric estimation of propensity score using series estimation can be seen as estimating the propensity score parametrically with a rich and flexible specification of the regressor vector. In such a context, Assumption 2.1 (ii) excludes cases with a number of series term comparable with the sample size. Note that the strict overlap assumption (Assumption 2.1 (iii)) is stronger than the usual one imposed for estimability of ATT parameter at $\sqrt{n}$-rate, which only assumes that $G(\cdot)$ is bounded away only from one. We assume $G(\cdot)$ bounded away also from zero by a technical reason; to ensure a uniform convergence property necessary in the subsequent local asymptotic analysis. Imposing the strict overlap assumption is standard in the literature, although the limited overlap can be a concern in empirical applications (see Crump, Hotz, Imbens, and Mitnik (2008) and Khan and Tamer (2010) for further discussion.)

Following the original analysis of FIC by Claeskens and Hjort (2003), we consider approximating the finite sample bias-variance trade-offs between parsimonious and complicated models by adopting

a local misspecification framework. For this purpose, let $\{P_n : n \in \mathbb{N}\} \subset \mathcal{P}$ be a drifting sequence of population distributions, which weakly converges to a fixed $P_0 \in \mathcal{P}$. Along $\{P_n : n \in \mathbb{N}\}$, let $\{\gamma_n = \gamma(P_n)\}$ be a sequence of the coefficient vector of W, where $\gamma(\cdot)$ is as defined in Assumption 2.1(ii). Under Assumption 2.1, the ATT parameter τ_n satisfies the following moment condition: at every n,

$$E_{P_n}\left[\frac{D_i Y_i}{Q_n} - \frac{G\left(W_i'\gamma_n\right)\left(1-D_i\right)Y_i}{Q_n\left(1-G\left(W_i'\gamma_n\right)\right)} - \tau_n\right] = 0,$$

where E_{P_n} is the expectation with respect to P_n, and $Q_n \equiv P_n\left(D=1\right)$.

Let $\hat{\gamma}$ be the maximum likelihood estimator for γ_n obtained from the parametric binary choice regression, and $\hat{Q} = \frac{1}{n}\sum_{i=1}^n D_i$. The *normalized propensity score weight (NPW) estimator* for ATT (Imbens (2004)) in the largest model estimates ATT by

$$\hat{\tau} = \frac{1}{n}\sum_{i=1}^n \left(\frac{D_i Y_i}{\frac{1}{n}\sum_{i=1}^n D_i} - \frac{G\left(W_i'\hat{\gamma}\right)\left(1-D_i\right)}{\left(1-G\left(W_i'\hat{\gamma}\right)\right)}Y_i \Bigg/ \frac{1}{n}\sum_{i=1}^n \frac{G\left(W_i'\hat{\gamma}\right)\left(1-D_i\right)}{\left(1-G\left(W_i'\hat{\gamma}\right)\right)}\right), \tag{2.1}$$

where the summation terms, $\frac{1}{n}\sum_{i=1}^n D_i$ and $\frac{1}{n}\sum_{i=1}^n \frac{G\left(W_i'\hat{\gamma}\right)\left(1-D_i\right)}{\left(1-G\left(W_i'\hat{\gamma}\right)\right)}$ that appear in the denominators are to guarantee that the weights for Y_i sum up to one. This normalization improves the asymptotic variance of the naive ATT estimator when a parametric propensity score estimate is used. The NPW estimator is simple to implement, and simulation evidence of Busso et al. (2011) suggests that its finite sample performance is excellent, compared to other estimators relying on the unconfoundedness assumption. Therefore, in what follows, we will restrict our focus to the NPW estimator.

In order for $\hat{\tau}$ to have an asymptotic distribution in the local asymptotics along $\{P_n\}$, we impose a set of regularity conditions on a set of data generating processes $\mathcal{P}$, which are collected in Assumption A.1 of Appendix A. In what follows, $T \xrightarrow{P_n} c$, or, equivalently, $T - c = o_{P_n}(1)$ means statistic T converges in probability to c along $\{P_n\}$, i.e., $\lim_{n\to\infty} P_n\left(|T-c| > \epsilon\right) = 0$ for any $\epsilon > 0$. We use $T \overset{P_n}{\rightsquigarrow} \mathcal{N}\left(\mu.\Sigma\right)$ to mean that statistic T converges in distribution to Gaussian along sampling sequence $\{P_n\}$.

Each candidate specification for the propensity score corresponds to a subvector of $W\left(X\right)$ used in the propensity score estimation. Let S index a selection of covariates of W to be used in the estimation of the first stage maximum likelihood for the propensity scores. The number of covariates included in specification S is denoted by $|S|$. We denote the set of candidate specifications by

$\mathcal{M}$. $\mathcal{M}$ does not have to exhaust all the possible subset vectors of W (X), and some regressors can be included in all the specifications in $\mathcal{M}$ if they are believed to be important in predicting one's treatment status. Let $\underline{S} = \cap\{S : S \in \mathcal{M}\}$ be the set of covariates that appear in every candidate model. The number of candidate models is denoted by $|\mathcal{M}|$. We assume that $|\mathcal{M}|$ is fixed and does not grow with the sample size. The subset of covariates to be excluded from S is indexed by its complement, S^c. Hence, $\underline{S}^c$ represents the set of covariates that can be excluded in some candidate model. $|S| \times 1$ subvectors of W and γ, corresponding to the selected covariates are denoted by W_S and γ_S, respectively. We define $|S| \times K$ matrix π_S such that pre-multiplying π_S to a $K \times 1$ vector yields the subvector corresponding to selection S,

$$\pi_S W = W_S, \;\; \pi_S \gamma = \gamma_S.$$

Given a selection of covariates S, let $\hat{\tau}_S$ be the NPW estimator for τ when W_S is included in the estimation of the parametric propensity score, i.e.,

$$\hat{\tau}_S = \frac{1}{n}\sum_{i=1}^{n}\left(\frac{D_i Y_i}{\frac{1}{n}\sum_{i=1}^{n} D_i} - \frac{G\left(W_{S,i}'\hat{\gamma}_S\right)(1-D_i)}{\left(1-G\left(W_{S,i}'\hat{\gamma}_S\right)\right)}Y_i \Bigg/ \frac{1}{n}\sum_{i=1}^{n}\frac{G\left(W_{S,i}'\hat{\gamma}_S\right)(1-D_i)}{\left(1-G\left(W_{S,i}'\hat{\gamma}_S\right)\right)}\right),$$

where $\hat{\gamma}_S$ is the maximum likelihood estimator for γ_S obtained in the first stage propensity score regression of D_i on $W_{S,i}$.[1]

Following the original analysis of FIC of Claeskens and Hjort (2003), we consider a sequence of DGPs, along which the largest model shrinks to the submodels in the following sense.

Assumption 2.2: *Along a sequence of data generating processes* $\{P_n\} \in \mathcal{P}$ converging weakly to $P_0 \in \mathcal{P}$, *the sequence of coefficients of* W, $\{\gamma_n \equiv \gamma(P_n)\}$, *converges to the benchmark value* $\gamma_0 \equiv \gamma(P_0) \in \mathbb{R}^K$ *at* $n^{-1/2}$*-rate*

$$\gamma_n - \gamma_0 = \frac{1}{\sqrt{n}}\delta, \qquad \delta \in \bar{\mathbb{R}}^K,$$

where we assume $\gamma_{0,\underline{S}^c} \equiv \pi_{\underline{S}^c}\gamma_0 = 0$, *i.e., the benchmark coefficients of the regressors that can be excluded in some candidate models are zeros.*

[1] As an alternative NPW estimator for ATT in model S, we may consider an overidentified GMM estimator for τ. Using the moment conditions $m_i(\theta)$ to be defined in Section 3 and an optimal choice of weighting matrix Ω, a GMM estimator for τ in model S minimizes $\left(\frac{1}{n}\sum m_i(\theta)\right)'\Omega\left(\frac{1}{n}\sum m_i(\theta)\right)$ subject to $\gamma_{S^c} = 0$. Although this GMM estimator has a smaller asymptotic variance than $\hat{\tau}_S$ considered in this paper, its computation is not as simple as $\hat{\tau}_S$. We therefore do not consider such overidentified GMM estimators in our analysis.

The localization parameter $\delta \in \bar{\mathbb{R}}^K$ measures the local deviation of the true coefficient values from the benchmark coefficient γ_0. The value of δ controls the asymptotic bias of the NPW estimator of each submodel, as will be shown below. Since the procedures to be proposed do not depend on the benchmark values for the coefficients that appear in every candidate model, we do not have to assume the value of $\gamma_{0,\underline{S}}$.

Let $E_{P_0}(\cdot)$ and $Var_{P_0}(\cdot)$ be the expectation and variance defined at probability law $P_0 \in \mathcal{P}$. In what follows, we use the following notation:

$$
\begin{aligned}
G &= G\left(W'\gamma_0\right), \quad g = g\left(W'\gamma_0\right) = \left.\frac{dG(z)}{dz}\right|_{z=W'\gamma_0}, \quad Q = P_0(D=1). \\
\mu_1(X) &= E_{P_0}[Y(1)|X], \quad \mu_0(X) = E_{P_0}[Y(0)|X], \quad \Delta\mu(X) = \mu_1(X) - \mu_0(X), \\
\alpha_0 &= E_{P_0}[Y(0)|D=1], \qquad \tau_0 = E_{P_0}[Y(1) - Y(0)|D=1], \\
\sigma_1^2(X) &= Var_{P_0}(Y(1)|X), \quad \sigma_0^2(X) = Var_{P_0}(Y(0)|X), \\
h &= \frac{D-G}{G(1-G)} gW,
\end{aligned}
$$

where $h \in \mathbb{R}^K$ is the $K \times 1$ score vector in the first stage maximum likelihood estimation for γ evaluated at $\gamma = \gamma_0$, i.e., $E_{P_0}(h) = 0$ holds.

In the next proposition, we derive the asymptotic distribution of each submodel NPW estimator along a DGP sequence of Assumption 2.2.

Proposition 2.1 *Suppose Assumption 2.1 and the regularity conditions stated in Assumption A.1 in Appendix A hold for a class of data generating processes $\mathcal{P}$, and Assumption 2.2 holds for a sequence of data generating processes $\{P_n\} \subset \mathcal{P}$. For each $S \in \mathcal{M}$, let h_S be a subvector of the score vector h defined by*

$$
h_S \equiv \pi_S h = \frac{(D - G(W'\gamma_0)) g(W'\gamma_0)}{G(W'\gamma_0)(1 - G(W'\gamma_0))} W_S.
$$

At the data generating process $P = P_0$, we define $L\{h_1|h_2\}$ as the linear projection of a random variable h_1 onto a random vector h_2 and $L^{\perp}\{h_1|h_2\}$ as its orthogonal complement, i.e., $L\{h_1|h_2\} = E_{P_0}(h_1 h_2') E_{P_0}(h_2 h_2')^{-1} h_2$ and $L^{\perp}\{h_1|h_2\} = h_1 - L\{h_1|h_2\}$.

The limiting distribution of $\hat{\tau}_S$ along $\{P_n\}$ converging weakly to P_0 is

$$
\sqrt{n}(\hat{\tau}_S - \tau_n) \overset{P_n}{\rightsquigarrow} \mathcal{N}\left(0, \omega_S^2\right) + bias_S(\delta),
$$

$$\omega_S^2 = SEB_{\tau,S} + \frac{1}{Q^2} E_{P_0}\left[L^{\perp}\left\{ (D-G)\left[\Delta\mu(X) - \tau_0 + \frac{1-2G}{1-G}\left(\mu_0(X) - \alpha_0\right)\right] \middle| h_S \right\}^2 \right], \tag{2.2}$$

$$bias_S(\delta) = \frac{1}{Q} E_{P_0}\left[L^{\perp}\left\{ \left(\frac{D-G}{1-G}\right)\left[\mu_0(X) - \alpha_0\right] \middle| h_S \right\} h'_{S^c} \right] \delta_{S^c}. \tag{2.3}$$

where $SEB_{\tau,S}$ is the semiparametrically efficient variance bound for τ under a priori restriction such that the propensity score is parametric and the relevant regressors are W_S, i.e., $P(D=1|X) = G(W_S'\gamma_S)$,

$$\begin{aligned} SEB_{\tau,S} = {} & E_{P_0}\left[\left(\frac{G}{Q}\right)^2 \left\{ \frac{\sigma_1^2(X)}{G} + \frac{\sigma_0^2(X)}{1-G} + \left(\Delta\mu(X) - \tau_0\right)^2 \right\} \right] \\ & + \frac{1}{Q^2} E_{P_0}\left[L\left\{ (D-G)\left[\Delta\mu(X) - \tau_0\right] | h_S \right\}^2 \right]. \end{aligned} \tag{2.4}$$

Proof. See Appendix A. ■

Before arguing analytical insights out of this proposition, it is worth clarifying the motivation of the local asymptotic analysis in the current context. The ultimate goal of the analysis is to obtain an estimator for the ATT that optimally balances out the finite sample variance-bias tradeoffs across small to large models. For this purpose, a sequence of DGPs as defined in Assumption 2.2 is used as a device for deriving a class of δ-indexed sampling distributions of the NPW estimators, in which the variance and bias of the estimators can be analyzed at the same stochastic order.[2] Hence, if a value of δ that can give accurate approximation of the actual sampling distributions of the NPW estimators were given, we could obtain an MSE minimizing optimal specification. However, a value of δ that gives accurate MSE approximation in a given situation remains unknown even in large n, so, unless one model dominates the others uniformly over δ, a data-driven selection of the optimal model involves the non-trivial step of handling uncertainty for δ, as discussed in Section 3.

Based on the analytical expressions for the variance and bias, the following remarks summarize some useful insights on the variance-bias tradeoffs in the ATT estimation.

Remark 2.1 *The variance expression of the submodel NPW estimator (2.2) consists of two terms that both depend on the selection of regressors. The first term corresponds to the semiparametric efficiency bound for τ constructed with a priori knowledge that the parametric propensity score*

[2] In contrast, if we consider a type of asymptotics where n increases to infinity with a fixed DGP, we would obtain any nonzero bias of a submodel estimator $\hat{\tau}_S$ to have stochastically larger order than the variance irrespective of the size of misspecification. This asymptotics may well provide a poor approximation for the finite sample MSEs, when a small model is misspecified only slightly.

depends only on the selected regressors. See Graham, de Xavier Pinto, and Egel (2012) for a derivation of this variance bound. This variance bound depends on S through the variance of the linear projection of $(D-G)\left[\Delta\mu\left(X\right)-\tau_0\right]$ *onto* h_S *the score vector of the parametric propensity score estimation with regressor vector* W_S*. The fact that the dimension of* h_S *equals to the dimension of* W_S *implies that* $SEB_{\tau,S}$ *weakly monotonically increases as more regressors are included in the propensity score, i.e.,* $SEB_{\tau,S} \leq SEB_{\tau,S'}$ *whenever* $S \subset S'$*.*

The second term of (2.2) captures the inefficiency of the NPW estimators relative to the semiparametric variance bound with the knowledge that model S is correctly specified. This relative inefficiency term is represented as the variance of the linear projection residuals of $(D-G)\left[\Delta\mu\left(X\right)-\tau_0+\frac{1-2G}{1-G}\left(\mu_0(X)-\alpha_0\right)\right]$ *onto* h_S*. Therefore, in contrast to* $SEB_{\tau,S}$*, the second term weakly decreases with the dimension of* W_S*.*

As a whole, whether having more regressors in the propensity score inflates the variance of $\hat{\tau}_S$ *or not depends on which of the two effects (inflation of* $SEB_{\tau,S}$ *versus the reduction of relative inefficiency) dominates. In the special case where the treatment effects are homogeneous, i.e.,*$\Delta\mu(X)=\tau_0$ *for all X, the first component in the variance expression* $SEB_{\tau,S}$ *no longer depends on S, so that adding more regressors never inflates the variance of the NPW estimator. In contrast, if treatment effects are heterogeneous, a smaller model can have an NPW estimator with a smaller variance than that of bigger models. For an example, see the Monte Carlo specification of Section 4.1.*

Remark 2.2 *The bias term shown in (2.3) is the inner product of the local misspecification parameter vector for excluded regressors* δ_{S^c} *and the correlation vector of* h_{S^c} *and the linear projection residual of* $\left(\frac{D-G}{1-G}\right)\left[\mu_0\left(X\right)-\alpha_0\right]$ *onto* h_S*. Clearly, the bias of a submodel NPW estimator is zero if* δ_{S^c} *is the zero vector. Even when* δ_{S^c} *is a nonzero vector, the bias of a submodel NPW estimator can become zero if these two vectors are orthogonal. This implies that, depending on the value of the local misspecification parameters, we can reduce the bias of a submodel NPW estimator by dropping some covariates that are useful to predict treatment status. Thus, there is no general monotonic relationship available between the squared bias and the number of included regressors.*

Remark 2.3 *As indicated by the relative inefficiency term appearing in (2.2), the NPW estimator is not semiparametrically efficient even when the propensity score specification in the submodel is correct. Recently, Graham et al. (2011) propose the Auxiliary-to-Study Tilting (AST) estimator for the ATT, that has a smaller asymptotic variance than the NPW estimator, and can achieve*

$SEB_{\tau,S}$ under the assumption that $\mu_1(X)$ and $\mu_0(X)$ are linear in a prespecified set of covariate vector used in the tilting step. The current local asymptotic analysis can be applied to the AST estimators, and the model selection and averaging for the AST estimators can be developed along the same line of analysis given in the next section. In this paper, however, we exclusively focus on the NPW estimator since the NPW estimator is easier to implement and numerically more stable than the AST estimator.[3]

3 Focussed Information Criterion and Model Averaging for ATT estimation

The analytical results of Section 2 show that, in the presence of treatment effect heterogeneity (i.e., $\Delta\mu(X)$ is not a constant), the MSEs of the NPW estimator approximated by the local asymptotics lead to nontrivial variance-bias tradeoffs between the small and large models, and an optimal selection of regressors that minimizes the MSE of $\hat{\tau}_S$ can be a proper subset of the regressors in the largest model. Since the localization parameter δ remains unknown even for large n, estimation of MSEs and selection of optimal specification crucially hinges on how the non-vanishing uncertainty of δ is dealt with.

In this section, we consider two ways estimate the MSEs: one is the unbiased estimation for MSEs along the same line as in the focussed information criterion (FIC) of Claeskens and Hjort (2003); another is a posterior estimation for MSEs that leads a model selection to a Bayes decision in the limit experiment. As a smoothed version of model selection and post-selection estimation, we subsequently consider how to optimally average the NPW estimators over the candidate models (model averaging).

3.1 MSE Minimizing Covariate Selection with FIC

Estimation of the MSE of each submodel NPW estimator is most easily formulated by focusing on the set of moment conditions that yields the NPW estimator in the largest model,

$$E_{P_n}\left[\mathbf{m}_i\left(\theta_n\right)\right] = 0,$$

[3]In the Monte Carlo study of Section 4, we tried to implement the covariate selection procedure based the AST estimator. In many Monte Carlo samples, however, we failed to compute the AST estimator apparently due to the lack of an interior solution in the tilting step. Busso et al. (2011) also remarks on such numerical instability of the AST estimator.

where $\theta_n = (\gamma_n', \alpha_n, \tau_n)'$ and

$$\mathbf{m}_i(\theta) = \begin{pmatrix} \frac{(D_i - G(W_i'\gamma))}{G(W_i'\gamma)[1-G(W_i'\gamma)]} g(W_i'\gamma) W_i \\ \left[D_i + (1-D_i)\left(\frac{G(W_i'\gamma)}{1-G(W_i'\gamma)}\right)\right](Y_i - \tau D_i - \alpha) \\ \left[D_i + (1-D_i)\left(\frac{G(W_i'\gamma)}{1-G(W_i'\gamma)}\right)\right](Y_i - \tau D_i - \alpha) D_i \end{pmatrix}.$$

These moment conditions are in the form of weighted least squares with weights $\left[D_i + (1-D_i)\left(\frac{G(W_i'\gamma)}{1-G(W_i'\gamma)}\right)\right]$, and have been used in Busso et al. (2011). Let

$$M = E_{P_0}\left[\frac{\partial}{\partial \theta'}\mathbf{m}_i(\theta)\Big|_{\theta=\theta_0}\right],$$
$$\Sigma = E_{P_0}\left[\mathbf{m}_i(\theta_0)\mathbf{m}_i(\theta_0)'\right],$$

which are consistently estimated by

$$\hat{M} = \frac{1}{n}\sum_{i=1}^{n}\frac{\partial}{\partial \theta'}\mathbf{m}_i\left(\hat{\theta}\right),$$
$$\hat{\Sigma} = \frac{1}{n}\sum_{i=1}^{n}\mathbf{m}_i\left(\hat{\theta}\right)\mathbf{m}_i\left(\hat{\theta}\right)',$$

where $\hat{\theta} = (\hat{\gamma}', \hat{\alpha}, \hat{\tau})$ is the estimator for θ in the largest model. See Lemma A.1 in Appendix A for these consistency claims. Using the selection matrix,

$$\underset{(|S|+2)\times(K+2)}{\Lambda_S} = \begin{pmatrix} \pi_S & & O \\ & 1 & \\ O & & 1 \end{pmatrix}$$

the asymptotic variance and the squared bias terms of $\sqrt{n}(\hat{\tau}_S - \tau_n)$ can be written as

$$\omega_S^2 = \text{the final element in the bottom row of} \quad (3.1)$$
$$\left(\Lambda_S M \Lambda_S'\right)^{-1}\Lambda_S \Sigma \Lambda_S'\left(\Lambda_S M' \Lambda_S'\right)^{-1},$$

$$bias_S^2(\delta) = \mathbf{b}_S'\delta_{S^c}\delta_{S^c}'\mathbf{b}_S, \quad (3.2)$$
$$\mathbf{b}_S' = \text{ the first } |S^c| \text{ elements of the row vector in the bottom row of}$$
$$\left(\Lambda_S M \Lambda_S'\right)^{-1}\Lambda_S M \Lambda_{S^c}'.$$

By plugging in $\hat{M}$ and $\hat{\Sigma}$, we obtain consistent estimators for ω_S^2 and $\mathbf{b}_S$, while the squared bias term involves the square of the local misspecification parameters $\delta_{S^c}\delta_{S^c}'$, for which a consistent estimator is not available.

The first method estimates MSEs by plugging in an asymptotically unbiased estimator for $\delta_{S^c}\delta_{S^c}'$. Note that

$$\sqrt{n}\left(\pi_{S^c}\hat{\gamma} - \gamma_{0,S^c}\right) \overset{P_n}{\rightsquigarrow} \mathcal{N}\left(\delta_{S^c}, \pi_{S^c}\mathcal{I}_\gamma^{-1}\pi_{S^c}'\right), \tag{3.3}$$

where $\mathcal{I}_\gamma = E_{P_0}\left(hh'\right)$ is the Fisher information for γ in the full model, which can be consistently estimated by

$$\hat{\mathcal{I}}_\gamma = \frac{1}{n}\sum_{i=1}^{n} \frac{g^2\left(W_i'\hat{\gamma}\right)}{G\left(W_i'\hat{\gamma}\right)\left[1 - G\left(W_i'\hat{\gamma}\right)\right]} W_i W_i'.$$

Since $\gamma_{0,S^c} = 0$ under Assumption 2.2, the maximum likelihood estimator for δ_{S^c} in the limit normal experiment of (3.3) is given by $\hat{\delta}_{S^c} = \sqrt{n}\pi_{S^c}\hat{\gamma}$. Since $E_{P_n}\left(\hat{\delta}_{S^c}\hat{\delta}_{S^c}'\right) \to \delta_{S^c}\delta_{S^c}' + \pi_{S^c}\mathcal{I}_\gamma^{-1}\pi_{S^c}'$ as $n \to \infty$, an asymptotically unbiased estimator for $\delta_{S^c}\delta_{S^c}'$ can be constructed as

$$\widehat{\delta_{S^c}\delta_{S^c}'} = \pi_{S^c}\left[n\hat{\gamma}\hat{\gamma}' - \hat{\mathcal{I}}_\gamma^{-1}\right]\pi_{S^c}'.$$

Based on this unbiased estimate of $\delta_{S^c}\delta_{S^c}'$, we estimate the squared bias term by

$$\widehat{bias_S^2}\left(\delta\right) = \max\left\{\hat{\mathbf{b}}_S'\pi_{S^c}\left[n\hat{\gamma}\hat{\gamma}' - \hat{\mathcal{I}}_\gamma^{-1}\right]\pi_{S^c}'\hat{\mathbf{b}}_S, 0\right\}, \tag{3.4}$$

where the max operators in these expressions are used to modify the negative estimates for the squared bias. With summing up the consistent variance estimator and the approximately unbiased estimator for the squared bias, we obtain FIC for specification S,

$$FIC(S) = \widehat{\omega_S^2} + \widehat{bias_S^2}\left(\delta\right).$$

The second method estimates MSEs based a posterior distribution for $\delta_{S^c}\delta_{S^c}'$ formed in the limit Gaussian experiment. This Bayesian way of making use of data is considered in Claeskens and Hjort (2003, Section 7), and the same idea can apply to the current context as well. Let $\mu\left(\delta_{\underline{S}^c}\right)$ be a prior distribution for $\delta_{\underline{S}^c} \in \mathbb{R}^{K-|\underline{S}^c|}$, that either reflects user's prior opinion about which $\delta_{\underline{S}^c}$ are more likely than others, or represents degrees of importance over the parameter space regaring the model selection performance. We update this prior based on the sufficient statistics for $\delta_{\underline{S}^c}$ in the

limit Gaussian experiment, $\hat{\delta}_{\underline{S}^c} = \sqrt{n}\pi_{\underline{S}^c}\hat{\gamma} \sim N(\delta_{\underline{S}^c}, \pi_{\underline{S}^c}\mathcal{I}_\gamma^{-1}\pi'_{\underline{S}^c})$. We then estimate the squared bias term of the NPW estimator in model S by plugging in $\hat{\mathbf{b}}_S$ and the posterior mean of $\delta_{S^c}\delta'_{S^c}$,

$$\begin{aligned} E_{\delta_{\underline{S}^c}|\hat{\delta}_{\underline{S}^c}}\left(bias_S^2(\delta)\right) &= \hat{\mathbf{b}}_S' E_{\delta_{\underline{S}^c}|\hat{\delta}_{\underline{S}^c}}\left(\delta_{S^c}\delta'_{S^c}\right)\hat{\mathbf{b}}_S \\ &= \hat{\mathbf{b}}_S'\pi_S\pi'_{\underline{S}^c} E_{\delta_{\underline{S}^c}|\hat{\delta}_{\underline{S}^c}}\left(\delta_{\underline{S}^c}\delta'_{\underline{S}^c}\right)\pi_{\underline{S}^c}\pi'_S\hat{\mathbf{b}}_S, \end{aligned}$$

where $E_{\delta_{\underline{S}^c}|\hat{\delta}_{\underline{S}^c}}(\cdot)$ is the expectation with respect to the posterior distribution of $\delta_{\underline{S}^c}$. A convenient specification for $\mu\left(\delta_{\underline{S}^c}\right)$ is a conjugate normal prior with mean ϕ and variance matrix Φ, with which the closed-form expression for $E_{\delta_{\underline{S}^c}|\hat{\delta}_{\underline{S}^c}}\left(bias_S^2(\delta)\right)$ is obtained as

$$E_{\delta_{\underline{S}^c}|\hat{\delta}_{\underline{S}^c}}\left(bias_S^2(\delta)\right) = \hat{\mathbf{b}}_S'\pi_{S^c}\pi'_{\underline{S}^c}\left(\overline{\delta}_{\underline{S}^c}\overline{\delta}'_{\underline{S}^c} + \left(\left(\pi_{\underline{S}^c}\hat{\mathcal{I}}_\gamma^{-1}\pi'_{\underline{S}^c}\right)^{-1} + \Phi^{-1}\right)^{-1}\right)\pi_{\underline{S}^c}\pi'_{S^c}\hat{\mathbf{b}}_S,$$

where $\overline{\delta}_{\underline{S}^c}$ is the posterior mean of $\delta_{\underline{S}^c}$, $\overline{\delta}_{\underline{S}^c} = \left(\left(\pi_{\underline{S}^c}\hat{\mathcal{I}}_\gamma^{-1}\pi'_{\underline{S}^c}\right)^{-1} + \Phi^{-1}\right)^{-1}\left(\left(\pi_{\underline{S}^c}\hat{\mathcal{I}}_\gamma^{-1}\pi'_{\underline{S}^c}\right)^{-1}\hat{\delta}_{\underline{S}^c} + \Phi^{-1}\phi\right)$. If the improper uniform distribution is used as a non-informative prior for $\delta_{\underline{S}^c}$, the posterior mean of the squared bias has the following simple form,

$$E_{\delta_{\underline{S}^c}|\hat{\delta}_{\underline{S}^c}}\left(bias_S^2(\delta)\right) = \hat{\mathbf{b}}_S'\pi_{S^c}\left(n\hat{\gamma}\hat{\gamma}' + \hat{\mathcal{I}}_\gamma^{-1}\right)\pi'_{S^c}\hat{\mathbf{b}}_S. \tag{3.5}$$

We refer to the MSE estimate formed via a posterior of $\delta_{\underline{S}^c}$ as Bayesian FIC (BFIC),

$$BFIC(S) = \widehat{\omega_S^2} + E_{\delta_{\underline{S}^c}|\hat{\delta}_{\underline{S}^c}}\left(bias_S^2(\delta)\right),$$

where $\widehat{\omega_S^2}$ is the consistent variance estimator as used in the construction of $FIC(S)$.[4]

The implementation of the covariate selection procedure is summarized as follows, regardless of whether the researcher decides to use FIC or BFIC:

FIC/BFIC based covariate selection for ATT estimation.

(Step 1) Specify a largest model (regressor vector W), in which Assumption 2.1 (ii) and (iii) are believed to be reasonable.

(Step 2) Run NPW estimation with the full regressor vector W, and, based on the parameter estimates in the largest model, obtain $\hat{M}$, $\hat{\Sigma}$, $\hat{\gamma}\hat{\gamma}'$, and $\hat{\mathcal{I}}_\gamma$.

[4]If the improper uniform prior for $\delta_{\underline{S}^c}$ is used, $E_{\delta_{\underline{S}^c}|\hat{\delta}_{\underline{S}^c}}\left(bias_S^2(\delta)\right) - \widehat{bias_S^2(\delta)} \geqq 0$ holds for any realization of $\hat{\delta}_{\underline{S}^c}$. This implies that, compared with FIC, BFIC (with the uniform prior) results inestimating the squared bias to be bigger.

(Step 3) Compute the FIC, or BFIC (with a give prior for $\delta_{\underline{S}^c}$) for each candidate model. Find the optimal selection of regressors, S, that minimizes $FIC(S)$ or $BFIC(S)$.

(Step 4) Using the optimal selection of regressors obtained in Step 3, we compute the NPW estimator for τ, and report it as a post-selection point estimator for τ.

3.2 MSE Minimizing Model Averaging

In this section, we consider an estimator for ATT in the following average form,

$$\hat{\tau}_{avg} = \sum_{S \in \mathcal{M}} c_S\left(\hat{\delta}_{\underline{S}^c}\right) \hat{\tau}_S, \tag{3.6}$$

where $c_S\left(\hat{\delta}_{\underline{S}^c}\right)$ are weights summing up to one, $\sum_{S\in\mathcal{M}} c_S\left(\hat{\delta}_{\underline{S}^c}\right) = 1$. $\hat{\tau}_{avg}$ is the weighted sum of the NPW estimators calculated in the candidate specifications, and the goal of model averaging is to choose the averaging weight so as to minimize the MSE of $\hat{\tau}_{avg}$.[5] The argument of $c_S\left(\hat{\delta}_{\underline{S}^c}\right)$ indicates that we specify weights to depend on data only through the sufficient statistics for the local misspecification parameter $\delta_{\underline{S}^c}$ in the limit experiment. Note that we allow some $c_S\left(\hat{\delta}_{\underline{S}^c}\right)$ to be negative since, by doing so, we can obtain optimal weights as an interior solution, whose closed form expression is available. In addition, we can potentially lower the asymptotic MSE of $\hat{\tau}_{avg}$ by relaxing the nonnegativity constraints of the weights.

Let $\hat{\mathbf{t}}_n$ be a $|\mathcal{M}| \times 1$ column vector consisting of $\{\sqrt{n}(\hat{\tau}_S - \tau_n) : S \in \mathcal{M}\}$. By noting that the bias expression obtained in (2.3) can be written as $\mathbf{b}_S'\pi_{S^c}\pi_{\underline{S}^c}'\delta_{\underline{S}^c}$, we can express the asymptotic distribution of $\left(\hat{\delta}_{\underline{S}^c}, \hat{\mathbf{t}}_n\right)$ as

$$\begin{pmatrix} \hat{\delta}_{\underline{S}^c} \\ \hat{\mathbf{t}}_n \end{pmatrix} \overset{P_n}{\rightsquigarrow} \mathcal{N}\left(\begin{pmatrix} \delta_{\underline{S}^c} \\ B\delta_{\underline{S}^c} \end{pmatrix}, \begin{pmatrix} \Omega_{11} & \Omega_{12} \\ \Omega_{21} & \Omega_{22} \end{pmatrix}\right), \tag{3.7}$$

where B is a $|\mathcal{M}| \times |\underline{S}^c|$ matrix, whose row vector corresponding to model S is $\mathbf{b}_S'\pi_{S^c}\pi_{\underline{S}^c}'$. The covariance matrix $\begin{pmatrix} \Omega_{11} & \Omega_{12} \\ \Omega_{21} & \Omega_{22} \end{pmatrix}$ is the limit covariance matrix of the moment conditions,

$$\begin{pmatrix} \begin{pmatrix} -\pi_{\underline{S}^c}\mathcal{I}_\gamma^{-1} & \mathbf{0} & \mathbf{0} \end{pmatrix} \\ T \end{pmatrix} \mathbf{m}_i(\theta_0),$$

[5] As an alternative class of averaging estimators for ATT, we can consider NPW estimators for ATT with averaged propensity scores plugged-in. Analysing optimal averaging weight for this class of ATT estimator is beyond the scope of this paper.

where T is a $|\mathcal{M}| \times (K+2)$ matrix with each row vector corresponding to model S being the bottom row vector of $-\left(\Lambda_S M \Lambda_S'\right)^{-1} \Lambda_S$. Note that Ω_{11} is a submatrix of the asymptotic variance matrix of the first stage MLE, $\Omega_{11} = \pi_{\underline{S}^c} \mathcal{I}_\gamma^{-1} \pi_{\underline{S}^c}'$. For a short-hand notation, denote averaging weights by $|\mathcal{M}| \times 1$ column vector $\mathbf{c}\left(\hat{\delta}_{\underline{S}^c}\right) = \left(c_S\left(\hat{\delta}_{\underline{S}^c}\right) : S \in \mathcal{M}\right)$. Given that $c_S\left(\hat{\delta}_{\underline{S}^c}\right)$ is continuous in $\hat{\delta}_{\underline{S}^c}$, the asymptotic MSE of $\sqrt{n}\left(\hat{\tau}_{avg} - \tau_n\right)$ under the local asymptotics is obtained as

$$MSE\left(\sqrt{n}\left(\hat{\tau}_{avg} - \tau_n\right)\right) \to E_{\hat{\delta}_{\underline{S}^c}|\delta_{\underline{S}^c}}\left[\mathbf{c}\left(\hat{\delta}_{\underline{S}^c}\right)' K\left(\hat{\delta}_{\underline{S}^c}, \delta_{\underline{S}^c}\right) \mathbf{c}\left(\hat{\delta}_{\underline{S}^c}\right)\right], \quad \text{as } n \to \infty, \tag{3.8}$$

where $E_{\hat{\delta}_{\underline{S}^c}|\delta_{\underline{S}^c}}$ is the expectation with respect to $\hat{\delta}_{\underline{S}^c} \sim \mathcal{N}\left(\delta_{\underline{S}^c}, \Omega_{11}\right)$, and $K\left(\hat{\delta}_{\underline{S}^c}, \delta_{\underline{S}^c}\right)$ is an $|\mathcal{M}| \times |\mathcal{M}|$ symmetric and positive-semidefinite matrix,

$$\begin{aligned} K\left(\hat{\delta}_{\underline{S}^c}, \delta_{\underline{S}^c}\right) &= \Omega_{22} - \Omega_{21}\Omega_{11}^{-1}\Omega_{12} \\ &\quad + \left(B - \Omega_{21}\Omega_{11}^{-1}\right)\left(\delta_{\underline{S}^c} - \hat{\delta}_{\underline{S}^c}\right)\left(\delta_{\underline{S}^c} - \hat{\delta}_{\underline{S}^c}\right)'\left(B - \Omega_{21}\Omega_{11}^{-1}\right)' \\ &\quad + \left(B - \Omega_{21}\Omega_{11}^{-1}\right)\left(\delta_{\underline{S}^c} - \hat{\delta}_{\underline{S}^c}\right)\hat{\delta}_{\underline{S}^c}' B' + B\hat{\delta}_{\underline{S}^c}\left(\delta_{\underline{S}^c} - \hat{\delta}_{\underline{S}^c}\right)'\left(B - \Omega_{21}\Omega_{11}^{-1}\right)' \\ &\quad + B\hat{\delta}_{\underline{S}^c}\hat{\delta}_{\underline{S}^c}' B'. \end{aligned}$$

See Appendix A for a derivation of (3.8).

By viewing the asymptotic MSE (3.8) as risk,the Bayes optimal averaging weights $\mathbf{c}^*(\cdot)$ with respect to prior $\mu\left(\delta_{\underline{S}^c}\right)$ is defined as,

$$\mathbf{c}^*(\cdot) = \arg \min_{\mathbf{c}(\cdot):\sum_S c_S(\cdot)=1} \int E_{\hat{\delta}_{\underline{S}^c}|\delta_{\underline{S}^c}}\left[\mathbf{c}\left(\hat{\delta}_{\underline{S}^c}\right)' K\left(\hat{\delta}_{\underline{S}^c}, \delta_{\underline{S}^c}\right) \mathbf{c}\left(\hat{\delta}_{\underline{S}^c}\right)\right] d\mu\left(\delta_{\underline{S}^c}\right). \tag{3.9}$$

We refer to the resulting averaging estimator as *BayesLE*. Since solving for the Bayes optimal weights is equivalent to minimizing the posterior risk, we obtain the following closed form expression of the Bayes optimal weights.

Proposition 3.1 *Let $\mu\left(\delta_{\underline{S}^c}\right)$ be a proper prior, and let $K_{post}\left(\hat{\delta}_{\underline{S}^c}\right)$ be the posterior expectation of $K\left(\hat{\delta}_{\underline{S}^c}, \delta_{\underline{S}^c}\right)$ given $\hat{\delta}_{\underline{S}^c} \sim \mathcal{N}(\delta_{\underline{S}^c}, \Omega_{11})$,*

$$K_{post}\left(\hat{\delta}_{\underline{S}^c}\right) \equiv E_{\delta_{\underline{S}^c}|\hat{\delta}_{\underline{S}^c}}\left[K\left(\hat{\delta}_{\underline{S}^c}, \delta_{\underline{S}^c}\right)\right]$$

If $K_{post}\left(\hat{\delta}_{\underline{S}^c}\right)$ is nonsingular almost surely in $\hat{\delta}_{\underline{S}^c}$, the Bayes optimal action for model averaging weight $\mathbf{c}^*\left(\hat{\delta}_{\underline{S}^c}\right)$ is unique almost surely in $\hat{\delta}_{\underline{S}^c}$, and is given by

$$\mathbf{c}^*\left(\hat{\delta}_{\underline{S}^c}\right) = \left[\mathbf{1}' K_{post}\left(\hat{\delta}_{\underline{S}^c}\right)^{-1} \mathbf{1}\right]^{-1} \left[K_{post}\left(\hat{\delta}_{\underline{S}^c}\right)^{-1} \mathbf{1}\right], \tag{3.10}$$

where $\mathbf{1}$ *is the vector of ones with length* $|\mathcal{M}|$.

Proof. See Appendix A. ■

If $\mu\left(\delta_{\underline{S}^c}\right)$ is specified to be conjugate normal with mean ϕ and variance Φ, then the conjugate normal posterior, $\delta_{\underline{S}^c}|\hat{\delta}_{\underline{S}^c} \sim \mathcal{N}\left(\overline{\delta}, \left(\Omega_{11}^{-1}+\Phi^{-1}\right)^{-1}\right)$, yields

$$\begin{aligned} K_{post}\left(\hat{\delta}_{\underline{S}^c}\right) &= \Omega_{22}-\Omega_{21}\Omega_{11}^{-1}\Omega_{12} \\ &+ \left[\left(B-\Omega_{21}\Omega_{11}^{-1}\right)\overline{\delta}_{\underline{S}^c}+\Omega_{21}\Omega_{11}^{-1}\hat{\delta}_{\underline{S}^c}\right]\left[\left(B-\Omega_{21}\Omega_{11}^{-1}\right)\overline{\delta}_{\underline{S}^c}+\Omega_{21}\Omega_{11}^{-1}\hat{\delta}_{\underline{S}^c}\right]' \\ &+ \left(B-\Omega_{21}\Omega_{11}^{-1}\right)\left(\Omega_{11}^{-1}+\Phi^{-1}\right)^{-1}\left(B-\Omega_{21}\Omega_{11}^{-1}\right)'. \end{aligned} \tag{3.11}$$

We accordingly obtain the optimal weights by plugging in consistent estimates of Ω's and B into $K_{post}\left(\hat{\delta}_{\underline{S}^c}\right)$, and apply formula (3.10).

Remark 3.1 *The main reason that Proposition 3.1 assumes proper* $\mu\left(\delta_{\underline{S}^c}\right)$ *is to guarantee that the Bayes risk (3.9) is bounded, and hence the minimization problem (3.9) has a well defined solution. In practice, however, forcing the researcher to have a proper prior may be restrictive if she/he does not have a credible prior opinion for* $\delta_{\underline{S}^c}$*, or she\he wishes to apply a non-informative prior for the purpose of reporting a default averaging estimate. If we specify* $\mu\left(\delta_{\underline{S}^c}\right)$ *to be uniform (Jeffreys' prior for the Gaussian means),* $K_{post}\left(\hat{\delta}_{\underline{S}^c}\right)$ *is well defined,*

$$K_{post}\left(\hat{\delta}_{\underline{S}^c}\right) = \Omega_{22}-\Omega_{21}\Omega_{11}^{-1}\Omega_{12}+\left(B-\Omega_{21}\Omega_{11}^{-1}\right)\Omega_{11}\left(B-\Omega_{21}\Omega_{11}^{-1}\right)'+B\hat{\delta}_{\underline{S}^c}\hat{\delta}_{\underline{S}^c}'B', \tag{3.12}$$

and the posterior risk has a well defined minimizer, as given by (3.10), despite that the resulting Bayes risk is unbounded.[6] *In Monte Carlo studies and an empirical application below, we examine performance of the averaging weights corresponding to the uniform prior.*

Remark 3.2 *Hjort and Claeskens (2003, Sec. 5.4) propose the following way of obtaining weights. Given localization parameter* $\delta_{\underline{S}^c}$ *and weight vector* $\mathbf{c}$*, the asymptotic MSE of* $\sqrt{n}\left(\hat{\tau}_{avg}-\tau_n\right)$ *is*

[6]One way to justify this averaging scheme would 1be to claim that the averaging weights corresponding to the uniform prior is obtained by a limit of the Bayes optimal weights with respect to a sequence of proper priors. Specifically, by noting that $K_{post}\left(\hat{\delta}_{\underline{S}^c}\right)$ of (3.11) converges to (3.12) as the prior variance matrix diverges to infinity, the averaging weights corresponding to the uniform prior can be obtained as the limit of the Bayes optimal weights along a sequence of conjugate priors with diverging prior variances.

written as $\mathbf{c}' E_{\hat{\delta}_{\underline{S}^c}|\delta_{\underline{S}^c}}\left[K\left(\hat{\delta}_{\underline{S}^c}, \delta_{\underline{S}^c}\right)\right]\mathbf{c} = \mathbf{c}'\left(\Omega_{22} - B\delta_{\underline{S}^c}\delta'_{\underline{S}^c}B'\right)\mathbf{c}$. *The weights proposed by Hjort and Claeskens minimize asymptotically unbiased estimator of this MSE,*

$$\mathbf{c}_{HC}\left(\hat{\delta}_{\underline{S}^c}\right) = \arg\min_{\mathbf{c}} \mathbf{c}'\left(\Omega_{22} - B\left(\hat{\delta}_{\underline{S}^c}\hat{\delta}'_{\underline{S}^c} - \Omega_{11}\right)B'\right)\mathbf{c},$$

which leads to

$$\mathbf{c}_{HC}\left(\hat{\delta}_{\underline{S}^c}\right) = \left[\mathbf{1}'\left(\Omega_{22} + B\left(\hat{\delta}_{\underline{S}^c}\hat{\delta}'_{\underline{S}^c} - \Omega_{11}\right)B'\right)^{-1}\mathbf{1}\right]^{-1}\left[\left(\Omega_{22} + B\left(\hat{\delta}_{\underline{S}^c}\hat{\delta}'_{\underline{S}^c} - \Omega_{11}\right)B'\right)^{-1}\mathbf{1}\right],$$

Note that $\mathbf{c}_{HC}\left(\hat{\delta}_{\underline{S}^c}\right)$ *can be shown to be different from the Bayes optimal weights of (3.11) for any of the conjugate normal priors, as well as the weights corresponding to the uniform prior presented in Remark 3.1.*

4 Monte Carlo Study

In this section, we perform a simulation experiment to study the behavior of the estimators discussed in Sections 2 and 3. We show that a bias-variance trade-off exists, and find MSE gains for the model averaging estimators proposed in Section 3. The results are especially favorable for the BayesLE estimator introduced in Proposition 3.1. We first consider a toy example that contains only one covariate in an oversimplified setting. Then, we consider a more realistic setting with an extensive set of simulations.

4.1 A Simplest Model

Suppose there is only one Bernoulli covariate $X \in \{0, 1\}$ with $P(X = 1) = p_X$ There are two candidate models. The large model has regressors $W = (1, X)'$ and the small model has only the intercept. In this simple setup, the propensity score can be specified as

$$P(D = 1|\, X) = G(W'\gamma) = \gamma^0 + \gamma^1 X.$$

We will introduce the local asymptotics by letting $\gamma^1 = \delta/\sqrt{n} \to 0$ as $n \to \infty$. At P_0, $\gamma^1 = 0$, we have $G(X) = \gamma^0$ so that D is independent of X. For simplicity, we normalize the regression functions of the potential outcomes by setting $\mu_1(1) = \mu_1(0) = \mu_0(1) = 0$, so that the only non-zero mean occurs for the control outcome with $X = 0$, i.e. $\mu_0(0) = -\mu$. In this simple setupthe average treatment effect for the treated is

$$\text{ATT} \quad = \quad \text{ATE}(1)P\left(X=1|D=1\right)+\text{ATE}(0)P\left(X=0|D=1\right)=-q_1\mu+0. \tag{4.1}$$

It is straightforward to show that the NPW estimator from the small model is

$$\hat{\tau}_{\emptyset}=\frac{1}{n_0}\sum_{i=1}^{n}\left(1-D_i\right)Y_i$$

with $n_0=\sum_i\left(1-D_i\right)$ and that the NPW estimator from the full model is

$$\hat{\tau}_f=-\hat{q}_1\hat{\mu}$$

where $\hat{\mu}$ is the sample analog of $\mu=E(Y|D=0,X=0)$, and $\hat{q}_1$ is the sample analog estimator for $q_1=P\left(X=1|D=1\right)$.

The previous expressions suggest that a bias-variance tradeoff may exist in this case. The small model estimator computes one sample analog using n_0 observations. The full model estimator uses a smaller number of observations for μ and requires the estimation of an additional term. To formalize this in a local asymptotic framework, we let

$$\gamma^1=P\left(D=1|X=1\right)-P\left(D=1|X=0\right)=\delta/\sqrt{n}$$

and refer the reader to Appendix B, where we show that the MSE in the small model minus the MSE in the large model is given by

$$\frac{2\gamma^0-1}{\gamma^0(1-\gamma^0)}p_X(1-p_X)\mu^2+\left(\frac{p_X(1-p_X)}{\gamma^0(1-\gamma^0)}\mu\delta\right)^2. \tag{4.2}$$

Hence, if $\gamma^0<1/2$, we can derive the range of $\delta=\sqrt{n}\gamma^1$, in which the small model has a smaller MSE:

$$-\sqrt{(1-2\gamma^0)\gamma^0(1-\gamma^0)}\leq\delta\leq\sqrt{(1-2\gamma^0)\gamma^0(1-\gamma^0)}.$$

Figure 1 presents the results for this case, assuming $p_X=1/2$ and $\gamma_0=0.3$. First, there is a clear bias-variance tradeoff which leads the small estimator to be preferred for a range of δ around 0. Second, the averaging estimator dominates the small estimator, and outperforms the full estimator over a larger range of δ, while the MSE improvement relative to the large model is not uniform over δ. The range of γ_1 for which the BayesLE estimator outperforms the full estimator is approximately $(-0.12,0.12)$, based on the simulations underlying Figure 1.

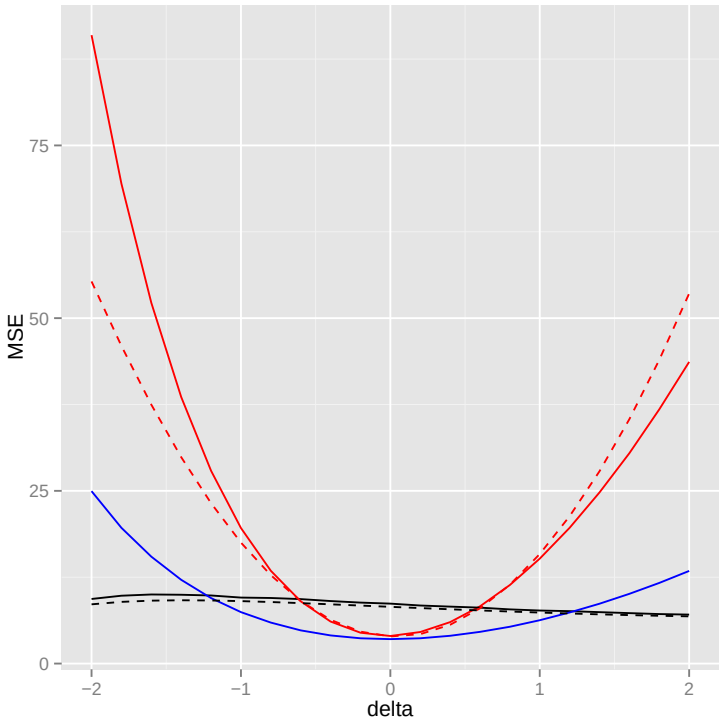

Figure 1: Analytical and simulations results for the simplest model with $p_X = q_0 = 1/2$, $n = 100$. The MSE is on the vertical axis, and δ is on the horizontal axis. Black: full model. Red: small model. Blue: BayesLE estimator. Dotted lines are based on analytic expressions, solid lines are based on simulations.

Parameter	Description	Value
n	Number of observations	300
K	Number of regressors	4
c	Correlation in covariates	0.7
$\gamma_1 = \beta_{11} = \beta_{01}$	Effect of X_1	1
$\alpha_0, \alpha_1, \alpha_D$	Constant terms	1; 2; 1
$\gamma_k = \beta_{1k}$, $k > 1$	Effect of X_k, $k > 1$ on $D, Y(1)$	0.1
β_{0k}, $k > 1$	Effect of X_k, $k > 1$ on $Y(0)$	0
$\sigma_0 = \sigma_1$	Conditional st. dev. $Y_i(0), Y_i(1)$	0.1

Table 1: Parameter values for the simulations in Section 4.2.

4.2 A More Realistic Model

For the purposes of this simulation study, we use the following model:

$$\begin{aligned} Y(0) &= \alpha_0 + X^{'}\beta_0 + u_0, \\ Y(1) &= \alpha_1 + X^{'}\beta_1 + u_1, \\ P(D = 1|\, X) &= \Lambda(\alpha_D + X^{'}\gamma), \\ X &\sim \mathcal{N}_K\left(0_K\,,\, cI_K + (1-c)\iota_K\iota_K^{'}\right), \end{aligned}$$

where Λ is the logistic function, and we assume that the error terms u_j are independent of the K explanatory variables X, and have a normal distribution with mean 0 and variances σ_j^2. The regressors follow a multivariate normal distribution, and the probability of treatment takes a logit form, assumed to be linear in the covariates. The parameter c controls the covariance structure of the regressors: $c = 1$ means that the regressors are independent; collinearity corresponds to $c = 0$. For the purpose of these simulations, we assume that the potential outcome equations are linear in the covariates. As can be seen from the list of benchmark parameter values in Table 4.2, we set the coefficients of X_1 in the propensity score equation and in the potential outcome equations equal to $\gamma_1 = \beta_{11} = \beta_{01} = 1$. We introduce treatment effect heterogeneity by setting the the coefficients for all other covariates equal to 0.1 for the treatment outcome, and 0 for the control outcome. Note that the first regressor (X_1) is very important, and should probably be included in estimation, but there may be a bias-variance tradeoff for the other covariates.

Given a number of regressors K, we either consider all $2^K - 1$ submodels, or the $2^{K-1} - 1$

submodels that include a constant term and the important regressor X_1. The model selection and model averaging estimators depend on estimators of the local misspecification parameter δ, and on the matrices B and Ω in equation (3.7). Estimators for δ and B are obtained from the full model, and the asymptotic covariance matrix Ω is estimated using the bootstrap (1000 bootstrap samples).

We will refer to the "full model" as the model with all regressors, and to the "small model" as the model that only includes X_1 and a constant term. On top of the submodel estimators, we study the following five estimators: (1) the "Best submodel" estimator, which is the submodel estimator with the lowest MSE across simulations; (2) the "BayesLE" estimator, described in Proposition 3.1; (3) the "HC" estimator, proposed by Hjort and Claeskens, described in Remark (3.2); (4) the "Selection" estimator, which chooses the estimator with the lowest estimated MSE; (5) the "invFIC" estimator, which weighs each submodel estimator by the inverse of its estimated MSE. The invFIC estimator is a naive weighting estimator that does not take into account the correlation between submodel estimators. Results for each model are based on 4000 replications.

The results for the benchmark simulations can be found in Table 4.2. We consider two scenarios: in the first one ("All submodels"), a researcher considers all submodels; in the second one ("Submodels with X_1"), the researcher has prior knowledge that the covariate X_1 is very important, and only considers models that include X_1. In the bottom row, we report the relative efficiency, which is defined as the MSE of the BayesLE estimator divided by the MSE of the estimator in the full model.

Several findings are worth noting. First, note that all the estimators that leave out the relevant regressor X_1 have poor performance due to omitted variable bias. Second, there is a clear bias-variance tradeoff: the small model (only X_1) outperforms the full model (all regressors). Third, the full model estimator has the lowest bias. Fourth, the BayesLE estimator seems to have the best overall performance, in terms of MSE. Finally, the performance of the selection procedure seems unaffected by the inclusion of poorly performing models (i.e. models without X_1), whereas including these poorly performing models improves the performance of the averaging estimators, and the BayesLE estimator in particular. This translates into the higher relative efficiency in the first scenario. In conclusion, the simulation results provide evidence in favor the BayesLE estimator, and suggest that it is robust against the inclusion of poorly performing models.

Sensitivity analysis. We now check whether the conclusions from the main results are robust to changes in the design parameters, and investigate the role of the design on the relative performance of the estimators. The results are summarized in Table 4.2. We consider seven designs. In

	All submodels			Submodels with X_1		
Estimator	Bias	Var	MSE	Bias	Var	MSE
$\{X_1\}$	2.13	3.07	3.11	2.29	3.04	3.06
$\{X_1, X_2\}$	2.05	3.18	3.22	2.23	3.18	3.23
$\{X_1, X_2, X_3\}$	1.94	3.33	3.37	2.19	3.27	3.32
$\{X_1, X_2, X_3, X_4\}$	1.94	3.40	3.43	2.22	3.36	3.40
$\{X_1, X_2, X_4\}$	2.04	3.26	3.30	2.26	3.28	3.33
$\{X_1, X_3\}$	2.02	3.24	3.28	2.25	3.12	3.17
$\{X_1, X_3, X_4\}$	2.03	3.30	3.34	2.29	3.22	3.28
$\{X_1, X_4\}$	2.13	3.16	3.20	2.33	3.13	3.18
$\{X_2\}$	80.10	1.46	65.61	-	-	-
$\{X_2, X_3\}$	75.22	1.55	58.13	-	-	-
$\{X_2, X_3, X_4\}$	72.07	1.60	53.55	-	-	-
$\{X_2, X_4\}$	75.20	1.52	58.06	-	-	-
$\{X_3\}$	80.09	1.48	65.62	-	-	-
$\{X_3, X_4\}$	75.24	1.53	58.13	-	-	-
$\{X_4\}$	80.06	1.45	65.54	-	-	-
Best submodel	2.13	3.08	3.11	2.29	3.04	3.06
Selection	3.18	3.06	3.16	3.46	3.01	3.13
BayesLE	3.90	2.09	2.24	4.56	2.87	3.08
HC	3.07	2.92	3.01	4.52	2.89	3.09
invFIC	4.67	2.61	2.83	2.45	3.11	3.17
Relative MSE Improvement		65%			90%	

Table 2: Simulation results for the benchmark setup. All values were multiplied by 100. Relative efficiency is the ratio of the BayesLE estimator's MSE to the MSE of the full model.

Design	(1) Smaller sample		(2) Larger sample		(3) Independent regressors		(4) Less relevant		(5) More relevant	
n	**150**		**1000**		300		300		300	
K	4		4		4		4		4	
c	0.7		0.7		**1**		0.7		0.7	
$\gamma_k, \beta_{1k}, \beta_{0k}, k > 1$	0.1		0.1		0.1		**0.05**		**0.2**	
σ_0, σ_1	0.1		0.1		0.1		0.1		0.1	
Estimator	Bias	MSE	Bias	MSE	Bias	MSE	Bias	MSE	Bias	MSE
Small model	4.27	5.13	0.64	1.06	1.63	2.54	1.97	2.83	2.63	3.71
Full model	4.67	5.91	0.54	1.16	1.51	2.82	1.93	3.10	2.52	4.52
Best submodel	4.27	5.13	0.64	1.06	1.63	2.54	1.97	2.83	2.63	3.71
Selection	6.11	5.51	1.32	1.05	2.69	2.54	3.07	2.86	4.18	4.03
BayesLE	5.53	4.25	2.49	0.76	1.71	2.17	3.94	2.05	4.04	2.75
HC	5.64	4.25	3.04	1.30	1.57	3.02	5.06	7.75	5.14	3.69
invFIC	8.88	4.97	1.65	0.97	3.77	2.30	4.40	2.54	5.84	3.48
Relative MSE Improvement	72%		65%		77%		66%		61%	

Table 3: Results of the sensivity analysis. Values are multiplied by 100.

design (1), we lower the sample size to 150. In design (2), we increase the sample size to 1000. In design (3), we consider a model with only two regressors. In design (4), we consider uncorrelated regressors. In design (5), we decrease the effect of the regressors in the selection and outcome equations by setting the coefficients to zero. In design 6, we increase the coefficients to 0.3, making the non-X_1 regressors more relevant.

The conclusions from the benchmark setup are unchanged. Some findings are worth pointing out. First, the lowest MSE for any design is achieved by the BayesLE estimator, except in design (3). Second, although the naive averaging estimator ("invFIC") provides an improvement over all submodel estimators and the selection estimator, it is generally outperformed by the BayesLE estimator.

Changes in the parameters of the design can have an impact on the absolute and relative performance of the estimators. First, it can be seen from the simulation output for designs (1) and (2) that increasing the sample size reduces the bias and variance for all estimators, without affecting the relative efficiency. Second, reducing the number of regressors in the full model leads to more precise estimators. There seems to be a lower relative efficiency from selection and averaging, as there is less of a bias-variance tradeoff. Third, reducing the correlation between the regressors reduces the variance of the estimates, but there seems to be no change in relative efficiency from model averaging. Fourth, increasing the importance of regressors $(X_2, \cdots, X_K)$ relative to X_1 reduces the precision of the estimators.

5 Empirical application

In this section, we apply the methods discussed in Sections 2 and 3 to the dataset analyzed in LaLonde (1986) and Dehejia and Wahba (1999). These papers estimate the impact of the National Supported Work Demonstration (NSW) on earnings. The NSW was implemented as a field experiment. Candidates were randomized across treatment and control groups. Those who were assigned to the treatment group benefitted from work experience, and some counselling. Due to the experimental implementation, the difference in post-intervention earnings of treatment and control groups is an unbiased estimator for the average effect of the NSW program on earnings. LaLonde shows that linear regression, fixed effects, and selection models fail to reproduce the experimental estimate, using as control group the members of the Panel Study on Income Dynamic (PSID) and the Current Population Survey (CPS). Dehejia and Wahba (DW) show that estimates obtained

Variable	Description	Always in?	Treated	CPS-1
age	Age (years)	Yes	25.82	33.23
education	Years of schooling	Yes	10.35	12.03
black	1 if black	Yes	0.84	0.07
re74	1974 earnings ($)	Yes	2096	14017
re75	1975 earnings ($)	Yes	1532	13651
hispanic	1 if hispanic	No	0.06	0.07
married	1 if married	No	0.19	0.71
nodegree	1 if no high school	No	0.71	0.30
age^2	(both or neither)	No	-	-
$re75^2$			-	-
Observations			185	15992

Table 4: Variables and transformation in our application. Column "Always in?" denotes whether we choose to include these covariates in the propensity score specification for each submodel. The last two columns report the sample means for the observations with $D_i = 1$ and $D_i = 0$, respectively.

using propensity score methods are closer to the experimental estimate.

A detailed description of the program and the data can be found in the aforementioned papers.[7] As in DW, we focus on the 185 observations on male participants in the treatment group for which pre-intervention incomes in both 1974 and 1975 are available. The non-experimental control group that we use is CPS-1.[8] Propensity score covariates and summary statistics are given in Table 4.

The experimental estimate for this subset is $1672 (standard error: $637), after a regression adjustment for age, education, and race.[9] Using stratification and matching on the estimated propensity score, DW's adjusted estimates are $1774 (standard error: $1152) and $1616 (standard error: $751), respectively. DW do not provide an in-depth discussion of how the covariates for the propensity score were chosen, but they describe that their results are sensitive to excluding higher

[7]The data is available from Raheev Dehejia's website. Last accessed: June 1, 2013. Location: http://users.nber.org/∼rdehejia/nswdata2.html.

[8]LaLonde (p. 611) provides details on the CPS-1 sample. We prefer the CPS over the PSID because of the larger sample size ($n = 15992$).

[9]The unadjusted estimate is $1794 with a standard error of $633.

	Full model		Small model	
Variable		SE	$\hat{\gamma}$	SE
age	72.66	2.93	−3.47	0.58
education	−0.99	0.55	−6.21	1.56
black	0.10	0.01	0.09	0.01
hispanic	0.03	0.01		
married	−1.58	0.26		
nodegree	0.29	0.06		
re74	−1.23	0.57	−0.90	0.54
re75	−5.99	1.24	−4.28	0.70
age^2	−18.25	2.31		
$re75^2$	0.88	0.48		
n		16177		16177

Table 5: Estimates and standard errors for the propensity score parameters in the full and small model. For the ease of comparison on the importance of each regressor, each coefficient estmate is multiplied by the standard deviation of the regressor.

order terms and to excluding 1974 earnings.[10]

We consider the set of variables and transformations in Table 4. Our choice of variables in the large model is identical to that in DW. The treatment and control groups have sizeable differences in terms of their observable characteristics, so a difference in means is unlikely to be unbiased for the average treatment effect. We consider 16 submodels: (i) for each variable in (hispanic, married, nodegree), we are unsure whether to include it in the propensity score or not; (ii) we are unsure whether to include squares (for re75 and age). We use a logit form for the selection equation. Table 5 presents the output for the propensity score estimation in the full and the small model. Clearly, omitting some of the covariates in the full model leads to biased estimation of γ, see for example the changes in the coefficient estimate for education. On the other hand, the coefficients are more

[10]DW use trimming in their empirical application, by discarding observations in the control group that have an estimated propensity that is lower than that the minimum estimated propensity score in the treatment group. We do not trim any observations in our empirical illustration. Note that whether trimming or not trimming any observations in the control group does not affect the estimand, as trimming is applied in the control group only, and the estimand is the ATT.

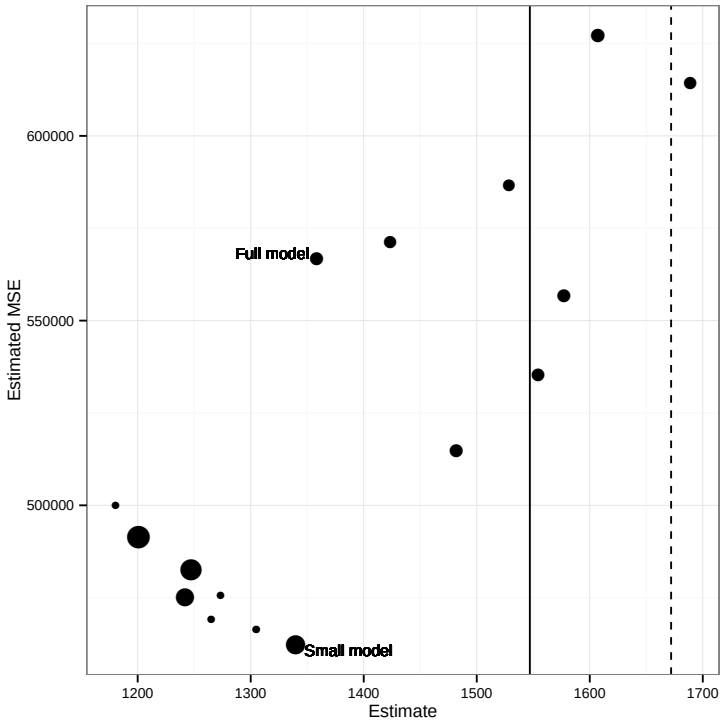

Figure 2: Estimates and their estimated mean squared errors. The vertical dotted line corresponds to the experimental estimate. The vertical solid line corresponds to the averaging estimate. For each submodel, the standard error is determined using 500 bootstrap replications. The radius of the circle is proportional to the weight assigned by the BayesLE estimator.

precisely estimated in the small model.

Figure 2 visualizes the NPW estimates and standard errors from all submodels. In terms of their standard errors, none of the submodel estimates gets close to the experimental estimate (horizontal dotted line). Some of the estimates are closer to the experimental estimate than the full model estimate, but this could be due to sampling error, as all the implied confidence intervals overlap.

Table 6 reports 95% confidence intervals for the experimental estimate, the full model estimate, and the BayesLE estimate, using the procedure described in Hjort and Claeskens (2008, p. 211). All confidence intervals are quite wide, which is consistent with the findings in Lalonde and DW. Note that the averaging procedure does not lead to more precise inference than using the full model. We want to stress that the objective of this paper is to come up with a point estimator that has good performance. The procedure is known to be conservative (Hjort and Claeskens, 2008, p. 211),

Method	Estimate	SE	95%-CI
Experimental	1672	637	[627, 2717]
Full model	1358	753	[123, 2593]
Bayes	1547	-	[-56, 2686]

Table 6: Estimates and confidence intervals for three procedures.

but addressing this issue is beyond the scope of this paper.

6 Concluding Remarks

We proposed covariate selection and model averaging procedures for propensity score weighted estimation of ATT by extending the framework of focussed information criterion and frequentist model averaging to the semiparametric estimation of ATT. The aim of these procedures is to construct a most accurate estimator for ATT in terms of MSE, provided that the unconfoundedness holds and propensity scores are correctly specified in a most complicated specification provided by the user. The resulting procedures are easy to implement, and can offer a reference estimate of the ATT that takes into account the uncertainty in propensity score specifications. Our Monte Carlo evidence shows that the proposed procedures, especially our model averaging scheme, significantly outperform the ATT estimator constructed in each candidate specification. We therefore recommend empirical researchers to report the model averaged estimate in the presence of specification uncertainty for propensity scores.

There are several issues and concerns that are not covered in the current framework of the analysis. First, the local asymptotic approximation becomes less precise as the number of regressors is large relative to the sample size, so that the proposed procedures will not be suitable to a situation, where the most complicated specification has too many regressors. Second, the normal approximation obtained via the local asymptotics will not be precise when the overlap condition is poorly satisfied. So, the performance of our selection/averaging procedure is questionable if the supports of the propensity scores have limited overlap, Third, this paper only concerns point estimation and does not consider precise post-selection/averaging inference for ATT. We leave these important issues for future research.

Appendix

A Proofs

We first introduce notations used throughout the appendix. For $P \in \mathcal{P}$, let $\tau(P) = E_P(Y(1) - Y(0)|D=1)$ and $\alpha(P) = E_P(Y(0)|D=1)$. Following Busso et al. (2011), we formulate the NPW estimation by the following system of just-identified moment conditions; ,

$$E_P[\mathbf{m}(Z_i, \theta(P))] = E_P\begin{pmatrix} \frac{(D_i - G(W_i'\gamma(P)))}{G(W_i'\gamma(P))[1-G(W_i'\gamma(P))]} g(W_i'\gamma(P)) W_i \\ \left[D_i + (1-D_i)\left(\frac{G(W_i'\gamma(P))}{1-G(W_i'\gamma(P))}\right)\right](Y_i - \tau(P) D_i - \alpha(P)) \\ \left[D_i + (1-D_i)\left(\frac{G(W_i'\gamma(P))}{1-G(W_i'\gamma(P))}\right)\right](Y_i - \tau(P) D_i - \alpha(P)) D_i \end{pmatrix} = 0. \quad \text{(A.1)}$$

where $Z_i = (Y_i, D_i, W(X_i))$ is a random vector of observation whose probability law is induced by $P \in \mathcal{P}$, and $\theta(P) = (\gamma(P)', \alpha(P), \tau(P))' \in R^{K+2}$ be a parameter vector solving these moment conditions when DGP is given at $P \in \mathcal{P}$. Note that the sample analogue of these moment conditions for the parameters yields the NPW estimator for τ in the largest model.

In what follows, we use the following notations. Let $\theta_n \equiv \theta(P_n) = (\gamma_n', \alpha_n, \tau_n)'$ and $\theta_0 \equiv \theta(P_0) = (\gamma_0', \alpha_0, \tau_0)'$. Let $\hat{\theta} = (\hat{\gamma}', \hat{\alpha}, \hat{\tau})'$ be the method of moment estimator in the largest model that solves $n^{-1}\sum_{i=1}^n m(Z_i, \theta) = 0$.

For each selection of covariates $S \in M$, we define

$$\begin{aligned} \gamma^S &= \pi_S'\pi_S\gamma + (I - \pi_S'\pi_S)\gamma_0 \\ \theta^S &= (\gamma^{S\prime}, \alpha, \tau)', \end{aligned} \quad \text{(A.2)}$$

where π_S is the selection matrix defined in the main text. γ^S is a $(K \times 1)$ vector obtained by replacing the elements of γ that are not selected in S with their benchmark values γ_0 (zeros by Assumption 2.2). In particular, for parameter sequence $\{\theta_n\}$ corresponding to $\{P_n\}$, we denote

$$\begin{aligned} \gamma_n^S &= \pi_S'\pi_S\gamma_n + (I - \pi_S'\pi_S)\gamma_0, \\ \theta_n^S &= (\gamma_n^{S\prime}, \alpha_n, \tau_n)'. \end{aligned}$$

For each $S \in \mathcal{M}$, define matrix

$$\underset{(|S|+2)\times K}{\Lambda_S} = \begin{pmatrix} \pi_S & & O \\ & 1 & \\ O & & 1 \end{pmatrix}.$$

Let $\hat{\gamma}_S$ be an $(|S| \times 1)$ vector of the MLE estimators obtained from the propensity score estimation with regressors W_S. Accordingly, we define a $(K \times 1)$ vector

$$\hat{\gamma}^S = \pi_S'\hat{\gamma}_S + (I - \pi_S'\pi_S)\gamma_0.$$

The NPW estimator in model S, interpreted as a method of moment estimator for θ in a reduced set of moment conditions, solves the following $(|S|+2)$-dimensional just-identifying sample moments,

$$\Lambda_S \left(\frac{1}{n} \sum_{i=1}^{n} \mathbf{m}\left(Z_i, \hat{\theta}^S\right) \right) = 0,$$

with

$$\hat{\theta}^S = \left(\hat{\gamma}^{S\prime}, \hat{\alpha}_S, \hat{\tau}_S\right)',$$

where $\hat{\tau}_S$ is the NPW ATT estimator in model S as defined in the main text, and $\hat{\alpha}_S$ is the corresponding value of α solving the sample moments of (A.1). Let $(\tilde{\gamma}_{n,S}, \tilde{\alpha}_{n,S}, \tilde{\tau}_{n,S})$ be the $(|S|+2)$-dimensional parameter vector that solves the population analogue of this reduced set of moment conditions for model S, i.e.,

$$E_{P_n}\left[\Lambda_S \mathbf{m}\left(Z_i, \tilde{\theta}_n^S\right)\right] = 0,$$

where

$$\tilde{\theta}_n^S = \left(\left(\pi_S' \tilde{\gamma}_{n,S} + \left(I - \pi_S' \pi_S\right)\gamma_0\right)', \tilde{\alpha}_{n,S}, \tilde{\tau}_{n,S}\right)'. \tag{A.3}$$

Note that $\tilde{\theta}_n^S$ generally differs from θ_n^S defined previously, while they will converge to the same limit θ_0 under the drifting sequence $\{P_n\}$ converging weakly to P_0, as will be shown in Lemma A.1.

Having introduced these notations, we now present a set of regularity conditions on a set of data generating processes $\mathcal{P}$, which $\{P_n\}$ and P_0 belong to.

Assumption A.1:

(i) The parameter space $\Theta = \left\{\theta(P) \in \mathbb{R}^{K+2} : P \in \mathcal{P}\right\}$ is compact.

(ii) There exists a compact set in R^K that contains the support of $W(X) \in \mathbb{R}^K$ for every $P \in \mathcal{P}$.

(iii) There exist $\lambda > 2$ such that $E_P\left[\left|Y(1) - E_P(Y(1)|X)\right|^\lambda | X\right] < \infty$ and $E_P\left[\left|Y(0) - E_P(Y(0)|X)\right|^\lambda | X\right] < \infty$ hold uniformly over $P \in \mathcal{P}$.

(iv) θ is identified uniformly over $\{P_n\} \subset \mathcal{P}$ in the sense that, for every $\epsilon > 0$, $0 = \|E_{P_n}[\mathbf{m}(Z, \theta_n)]\| < \inf_{\{\theta : \|\theta - \theta_n\| \geq \epsilon\}} \|E_{P_n}[\mathbf{m}(Z, \theta)]\|$ holds for all P_n.

(v) In each model $S \in \mathcal{M}$, $\tilde{\theta}_n^S$ defined in (A.3) is identified uniformly over $\{P_n\}$ in the sense that for every $\epsilon > 0$, $0 = \left\|E_{P_n}\left[\Lambda_S \mathbf{m}\left(Z, \tilde{\theta}_n^S\right)\right]\right\| < \inf_{\{\theta^S : \|\theta^S - \theta_n\| \geq \epsilon\}} \left\|E_{P_n}\left[\Lambda_S \mathbf{m}\left(Z, \theta^S\right)\right]\right\|$ holds for all P_n, where θ^S is as defined in (A.2).

The following lemmas are used to prove Proposition 2.1 and the consistency claims made in Section 3 of the main text.

Lemma A.1 *Suppose $\mathcal{P}$ satisfies Assumption 2.1 and Assumption A.1. Let $\{P_n\} \in \mathcal{P}$ be a sequence of data generating process converging weakly to $P_0 \in \mathcal{P}$.*

(i) $\left\|\hat{\theta} - \theta_n\right\| = o_{P_n}(1)$.

(ii) *In addition, assume Assumption 2.2.* $\left\|\hat{\theta}^S - \theta_n\right\| = o_{P_n}(1)$ *for every* $S \in M$.

(iii) *Let* $M = E_{P_0}\left[\frac{\partial}{\partial \theta'}\mathbf{m}(Z_i, \theta_0)\right]$.

$$\frac{1}{n}\sum_{i=1}^{n}\frac{\partial}{\partial \theta'}\mathbf{m}\left(Z_i, \hat{\theta}\right) - M = o_{P_n}(1),$$
$$\frac{1}{n}\sum_{i=1}^{n}\frac{\partial}{\partial \theta'}\mathbf{m}(Z_i, \theta_n) - M = o_{P_n}(1).$$

(iv) *Let* $\Sigma = E_{P_0}\left[\mathbf{m}(Z_i, \theta_0)\mathbf{m}(Z_i, \theta_0)'\right]$ *and* $\Sigma_n = E_{P_n}\left[\mathbf{m}(Z, \theta_n)\mathbf{m}(Z, \theta_n)'\right]$.

$$\frac{1}{n}\sum_{i=1}^{n}\mathbf{m}\left(Z_i, \hat{\theta}\right)\mathbf{m}\left(Z_i, \hat{\theta}\right)' - \Sigma = o_{P_n}(1),$$
$$\Sigma_n - \Sigma = o(1).$$

(v) $\frac{1}{\sqrt{n}}\sum_{i=1}^{n}\mathbf{m}(Z_i, \theta_n) \overset{P_n}{\rightsquigarrow} \mathcal{N}(0, \Sigma)$.

Proof. Let $\mathbf{m}_n(\theta) = \frac{1}{n}\sum_{i=1}^{n}\mathbf{m}(Z_i, \theta)$. To prove (i), we first show that, under Assumption 2.1 and A.1, uniform weak consistency of the moment conditions along $\{P_n\}$ holds, i.e., $\sup_\theta \|\mathbf{m}_n(\theta) - E_{P_n}[\mathbf{m}(Z, \theta)]\| = o_{Pn}(1)$. Let $\mathcal{F} \equiv \{\mathbf{m}(\cdot, \theta) : \theta \in \Theta\}$ be the class of moment functions indexed by θ. Under the overlapping assumption (Assumptions 2.1(iii)) and A.1 (i) and (ii), the moment conditions are Lipshitz continuous with bounded variation $\tilde{F}(Z)$,

$$\left\|\mathbf{m}(Z, \theta) - \mathbf{m}\left(Z, \tilde{\theta}\right)\right\| \leq \tilde{F}(Z)\left\|\theta - \tilde{\theta}\right\| \quad \text{for all } \theta, \tilde{\theta} \in \Theta,$$

where $\tilde{F}(Z) = \max\{c_1 W' W, c_2 |Y|, 1\}$ with some positive universal constants $c_1 < \infty$ and $c_2 < \infty$. Then, the covering number of the class of functions, $F = \{\|\mathbf{m}(Z, \theta) - \mathbf{m}(Z, \theta^*)\| : \theta \in \Theta\}$ for a fixed $\theta^* \in \Theta$, is bounded above by

$$N\left(\epsilon\left\|\tilde{F}\right\|, \mathcal{F}, \|\cdot\|\right) \leq \left[\frac{2diam(\Theta)}{\epsilon}\right]^{K+2},$$

for arbitrary seminorm $\|\cdot\|$ defined on $\mathcal{F}$ (Theorem 2.7.11 of van der Vaart and Wellner (1996)). Note that $\mathcal{F}$ has envelope $F(Z) = \tilde{F}(Z)\, diam(\Theta)$, and the covering number of $\mathcal{F}$ when the radius is set at $\epsilon\|F\|$ satisfies

$$N(\epsilon\|F\|, \mathcal{F}, \|\cdot\|) \leq \left[\frac{2}{\epsilon}\right]^{K+2} < \infty$$

for all $\epsilon > 0$. This leads to the bounded entropy number condition for $\mathcal{F}$. Furthermore, Assumption A.1 (ii) and (iii) assure the envelope is integrable uniformly over $\mathcal{P}$,$E_P[F(Z)] < \infty$ for all $P \in \mathcal{P}$. Hence, by Theorem 2.8.1 of van der Vaart and Wellner (1996), we obtain, for every $\{P_n\} \in \mathcal{P}$,

$$\sup_{\theta} \|\mathbf{m}_n(\theta) - E_{P_n}[\mathbf{m}(Z,\theta)]\| = o_{Pn}(1). \tag{A.4}$$

The definion of $\hat{\theta}$ and the uniform convergence of the moment conditions implies

$$0 = \left\|\mathbf{m}_n\left(\hat{\theta}\right)\right\| \leq \|\mathbf{m}_n(\theta_n)\| = \|E_{P_n}[\mathbf{m}(Z,\theta_n)]\| + o_{P_n}(1).$$

Accordingly, by noting $\|E_{P_n}[\mathbf{m}(Z,\theta_n)]\| - \left\|E_{P_n}\left[\mathbf{m}\left(Z,\hat{\theta}\right)\right]\right\| \leq 0$, we have

$$\left\|\mathbf{m}_n\left(\hat{\theta}\right)\right\| - \left\|E_{P_n}\left[\mathbf{m}\left(Z,\hat{\theta}\right)\right]\right\| \leq \|E_{P_n}[\mathbf{m}(Z,\theta_n)]\| - \left\|E_{P_n}\left[\mathbf{m}\left(Z,\hat{\theta}\right)\right]\right\| + o_{P_n}(1) \leq o_{P_n}(1).$$

Since, the left hand side quantity is $o_{P_n}(1)$ by (A.4), we obtain

$$\|E_{P_n}[\mathbf{m}(Z,\theta_n)]\| - \left\|E_{P_n}\left[\mathbf{m}\left(Z,\hat{\theta}\right)\right]\right\| = o_{P_n}(1) \tag{A.5}$$

By Assumption A.1(iv), for every $\epsilon > 0$, there exists $\lambda > 0$, such that $\|E_{P_n}[\mathbf{m}(Z,\theta_n)]\| - \|E_{P_n}[\mathbf{m}(Z,\theta)]\| < -\lambda$ holds for all P_n and $\|\theta - \theta_n\| \geq \epsilon$. Hence, we have

$$P_n(\left\|\hat{\theta} - \theta_n\right\| \geq \epsilon) \leqq P_n\left(\|E_{P_n}[\mathbf{m}(Z,\theta_n)]\| - \left\|E_{P_n}\left[\mathbf{m}\left(Z,\hat{\theta}\right)\right]\right\| < -\lambda\right),$$

where the right hand side converges to zero by (A.5). This leads to $\left\|\hat{\theta} - \theta_n\right\| = o_{P_n}(1)$.

To show (ii), consider

$$\left\|\hat{\theta}^S - \theta_n\right\| \leq \left\|\hat{\theta}^S - \tilde{\theta}_n^S\right\| + \left\|\tilde{\theta}_n^S - \theta_n^S\right\| + \left\|\theta_n^S - \theta_n\right\|. \tag{A.6}$$

In what follows, we prove each term in the right hand side vanishes asymptotically.

By a procedure similar to the proof of (i), we have

$$0 = \left\|\Lambda_S \mathbf{m}_n\left(\hat{\theta}^S\right)\right\| \leq \left\|\Lambda_S \mathbf{m}_n\left(\tilde{\theta}_n^S\right)\right\| = \left\|E_{P_n}\left[\Lambda_S \mathbf{m}\left(Z_i,\tilde{\theta}_n^S\right)\right]\right\| + o_{P_n}(1),$$

where the second equality follows from the uniform convergence of (A.4). Therefore,

$$\left\|\Lambda_S \mathbf{m}_n\left(\hat{\theta}^S\right)\right\| - \left\|E_{P_n}\left[\Lambda_S \mathbf{m}\left(Z_i,\hat{\theta}^S\right)\right]\right\| \leq \left\|E_{P_n}\left[\Lambda_S \mathbf{m}\left(Z_i,\tilde{\theta}_n^S\right)\right]\right\| - \left\|E_{P_n}\left[\Lambda_S \mathbf{m}\left(Z_i,\hat{\theta}^S\right)\right]\right\| + o_{P_n}(1) \leq o_{P_n}(1).$$

By repeating the same argument as in the proof of (i), Assumption A.1(v) and (A.5) yield $\left\|\hat{\theta}^S - \tilde{\theta}_n^S\right\| = o_{P_n}(1)$. Assumption 2.2 implies that the third term in the right hand side of (A.6) is $o(1)$, because $\left\|\theta_n^S - \theta_n\right\| \leq \|\gamma_n - \gamma_0\| = O\left(n^{-1/2}\right)$. In order to show that the second term in the right hand side of (A.6) is $o(1)$, consider applying the mean value expansion to $\mathbf{m}\left(Z,\theta_n^S\right)$ and take the expectation $E_{P_n}(\cdot)$,

$$\begin{aligned} E_{P_n}\left[\Lambda_S \mathbf{m}\left(Z,\theta_n^S\right)\right] &= E_{P_n}\left[\Lambda_S \mathbf{m}(Z,\theta_n)\right] + E_{P_n}\left[\Lambda_S \frac{\partial}{\partial \theta'}\mathbf{m}\left(Z,\bar{\theta}\right)\right]\left(\theta_n^S - \theta_n\right) \\ &= \Lambda_S E_{P_n}\left[\frac{\partial}{\partial \theta'}\mathbf{m}\left(Z,\bar{\theta}\right)\right]\left(\theta_n^S - \theta_n\right), \end{aligned}$$

Since $E_{P_n}\left[\Lambda_S \frac{\partial}{\partial\theta'}\mathbf{m}\left(Z,\bar{\theta}\right)\right] < \infty$ by Assumptions 2.1 (iii) and A.1 (ii), and $\left\|\theta_n^S - \theta_n\right\| = o(1)$ as shown above, we have $E_{P_n}\left[\Lambda_S \mathbf{m}\left(Z,\theta_n^S\right)\right] = o(1)$. Accordingly, we have

$$\left\|E_{P_n}\left[\Lambda_S \mathbf{m}\left(Z,\tilde{\theta}_n^S\right)\right]\right\| - \left\|E_{P_n}\left[\Lambda_S \mathbf{m}\left(Z,\theta_n^S\right)\right]\right\| = o(1).$$

The identification assumption (Assumption A.1(v)) then implies $\left\|\tilde{\theta}_n^S - \theta_n^S\right\| = o(1)$. Thus, we have shown that the right hand side of (A.6) is $o_{P_n}(1)$.

To show (iii), apply the triangular inequality to $\left(\frac{1}{n}\sum_{i=1}^n \frac{\partial}{\partial\theta'}\mathbf{m}\left(Z_i,\theta\right)\right)_{kl}$ where $(A)_{kl}$ means (k,l)-element of matrix A.

$$\begin{aligned}
&\left|\left(\frac{1}{n}\sum_{i=1}^n \frac{\partial}{\partial\theta'}\mathbf{m}\left(Z_i,\hat{\theta}\right) - E_{P_0}\left(\frac{\partial}{\partial\theta'}\mathbf{m}\left(Z_i,\theta_0\right)\right)\right)_{kl}\right| \\
\leq &\left|\left(\frac{1}{n}\sum_{i=1}^n \frac{\partial}{\partial\theta'}\mathbf{m}\left(Z_i,\hat{\theta}\right) - \frac{1}{n}\sum_{i=1}^n \frac{\partial}{\partial\theta'}\mathbf{m}\left(Z_i,\theta_0\right)\right)_{kl}\right| \\
&+\left|\left(\frac{1}{n}\sum_{i=1}^n \frac{\partial}{\partial\theta'}\mathbf{m}\left(Z_i,\theta_0\right) - E_{P_n}\left(\frac{\partial}{\partial\theta'}\mathbf{m}\left(Z_i,\theta_0\right)\right)\right)_{kl}\right| \\
&+\left|\left(E_{P_n}\left(\frac{\partial}{\partial\theta'}\mathbf{m}\left(Z_i,\theta_0\right)\right) - E_{P_0}\left(\frac{\partial}{\partial\theta'}\mathbf{m}\left(Z_i,\theta_0\right)\right)\right)_{kl}\right|.
\end{aligned} \tag{A.7}$$

Note that Assumption A.1 (i) and (ii) imply that every element of the derivative matrix $\frac{\partial}{\partial\theta'}\mathbf{m}\left(Z,\theta\right)$ is Lipshitz continuous in θ,

$$\left|\left(\frac{\partial}{\partial\theta'}\mathbf{m}\left(Z,\theta\right) - \frac{\partial}{\partial\theta'}\mathbf{m}\left(Z,\tilde{\theta}\right)\right)_{ij}\right| \leq f_{kl}\left(Z\right)\left\|\theta - \tilde{\theta}\right\| \quad \text{for all } \theta,\tilde{\theta}\in\Theta,$$

with $E_P\left(f_{ij}\left(Z\right)\right) < \infty$ for all $P \in \mathcal{P}$. Hence, the first term of (A.7) is bounded by

$$\left|\left(\frac{1}{n}\sum_{i=1}^n \frac{\partial}{\partial\theta'}\mathbf{m}\left(Z_i,\hat{\theta}\right) - \frac{1}{n}\sum_{i=1}^n \frac{\partial}{\partial\theta'}\mathbf{m}\left(Z_i,\theta_0\right)\right)_{kl}\right| \leq \left(\frac{1}{n}\sum_{i=1}^n f_{kl}\left(Z_i\right)\right)\left\|\hat{\theta} - \theta_0\right\| = o_{P_n}(1),$$

where $\left\|\hat{\theta} - \theta_0\right\| \leq \left\|\hat{\theta} - \theta_n\right\| + \|\theta_n - \theta_0\| = o_{P_n}(1)$ by Lemma A.1 (i) and Assymption 2.2, and $\left(\frac{1}{n}\sum_{i=1}^n f_{kl}\left(Z_i\right)\right) = E_{P_n}\left(f\left(Z_i\right)\right) + o_{P_n}(1)$ by the weak law of large number for triangular arrays (e.g. Lemma 11.4.2 of Lehman and Romano (2005)). Similarly, by the weak law of large number for triangular arrays, the second term in the right-hand side of (A.7) is $o_{P_n}(1)$ as well. The final term in the right-hand side of (A.7) is $o(1)$ since P_n converges weakly to P_0. Hence, we obtain $\frac{1}{n}\sum_{i=1}^n \frac{\partial}{\partial\theta'}\mathbf{m}\left(Z_i,\hat{\theta}\right) \overset{P_n}{\to} E_{P_0}\left(\frac{\partial}{\partial\theta'}\mathbf{m}\left(Z_i,\theta_0\right)\right)$. If we replace $\hat{\theta}$ in (A.7) with θ_n, the same argument can apply, and yields the desired conclusion.

A proof of (iv) can be obtained by repeating the same argument as in the proof of Lemma A.1(iii), so we omit the proof of (iv) for brevity.

To show (v), note that the strict overlap assumption and Assumption A.1(ii) and (iii) assure that the Lindberg condition holds for each moment condition in $\mathbf{m}\left(Z_i,\theta\right)$. Therefore, the Lindberg-Feller central limit theorem for triangular arrays leads to

$$\Sigma_n^{-1/2}\left(\frac{1}{\sqrt{n}}\sum_{i=1}^n \mathbf{m}\left(Z_i,\theta_n\right)\right) \overset{P_n}{\rightsquigarrow} \mathcal{N}\left(0,I_{K+2}\right).$$

By the assertion of Lemma A.1 (iv), $\Sigma^{1/2}\Sigma_n^{-1/2} \to I_{K+2}$, so

$$\begin{aligned}\frac{1}{\sqrt{n}}\sum_{i=1}^{n}\mathbf{m}\left(Z_i,\theta_n\right) &= \Sigma^{1/2}\Sigma_n^{-1/2}\left(\frac{1}{\sqrt{n}}\sum_{i=1}^{n}\mathbf{m}\left(Z_i,\theta_n\right)\right)+o_{P_n}\left(1\right)\\ &\overset{P_n}{\rightsquigarrow}\mathcal{N}\left(0,\Sigma\right).\end{aligned}$$

■

Proof of Proposition 2.1. The sample moment conditions that yield the NPW estimator in submodel S can be written as

$$0=\Lambda_S\mathbf{m}_n\left(\hat{\theta}^S\right).$$

By the mean value expansion around θ_n, we have

$$\begin{aligned}0&=\Lambda_S\mathbf{m}_n\left(\theta_n\right)+\Lambda_S\left[\frac{\partial}{\partial\theta'}\mathbf{m}_n\left(\bar{\theta}\right)\right]\begin{pmatrix}\hat{\gamma}^S-\gamma_n\\ \hat{\alpha}_S-\alpha_n\\ \hat{\tau}_S-\tau_n\end{pmatrix}\\ &=\Lambda_S\mathbf{m}_n\left(\theta_n\right)+\Lambda_S\left[\frac{\partial}{\partial\theta'}\mathbf{m}_n\left(\bar{\theta}\right)\right]\left[\Lambda_S'\begin{pmatrix}\hat{\gamma}_S-\gamma_{n,S}\\ \hat{\alpha}_S-\alpha_n\\ \hat{\tau}_S-\tau_n\end{pmatrix}-\Lambda_{S^c}'\begin{pmatrix}\gamma_{n,S^c}-\gamma_{0,S^c}\\ 0\\ 0\end{pmatrix}\right],\end{aligned}$$

where $\bar{\theta}$ is a convex combination of $\hat{\theta}^S$ and θ_n. Here, the second equality is obtained by plugging in $\hat{\gamma}^S=\pi_S'\hat{\gamma}_S+\pi_{S^c}'\gamma_0$. By Lemma A.1(ii), $\left\|\bar{\theta}-\theta_n\right\|=o_{P_n}\left(1\right)$. Then, Lemma A.1(iii) ensures $\frac{\partial}{\partial\theta'}\mathbf{m}_n\left(\bar{\theta}\right)-M=o_{P_n}\left(1\right)$. Therefore, by Lemma A.1(v) and Assumption 2.2, the asymptotic distribution of $\begin{pmatrix}\hat{\gamma}_S-\gamma_{n,S}\\ \hat{\alpha}_S-\alpha_n\\ \hat{\tau}_S-\tau_n\end{pmatrix}$ is obtained as

$$\begin{aligned}\sqrt{n}\begin{pmatrix}\hat{\gamma}_S-\gamma_{n,S}\\ \hat{\alpha}_S-\alpha_n\\ \hat{\tau}_S-\tau_n\end{pmatrix} &= -\left(\Lambda_S M\Lambda_S'\right)^{-1}\Lambda_S\left(\sqrt{n}\mathbf{m}_n\left(\theta_n\right)\right)+\left(\Lambda_S M\Lambda_S'\right)^{-1}\Lambda_S M\Lambda_{S^c}'\begin{pmatrix}\delta_{S^c}\\ 0\\ 0\end{pmatrix}+o_{P_n}\left(1\right).\\ &\overset{P_n}{\rightsquigarrow}-\left(\Lambda_S M\Lambda_S'\right)^{-1}\Lambda_S\times\mathcal{N}\left(0,\Sigma\right)+\left(\Lambda_S M\Lambda_S'\right)^{-1}\Lambda_S M\Lambda_{S^c}'\begin{pmatrix}\delta_{S^c}\\ 0\\ 0\end{pmatrix}\end{aligned}\tag{A.8}$$

In order to compute the asymptotic variance of $\sqrt{n}\left(\hat{\tau}_S-\tau_n\right)$, we focus on the variance of the bottom element of $-\left(\Lambda_S M\Lambda_S'\right)^{-1}\Lambda_S\mathbf{m}\left(Z_i,\theta_0\right)$ with $Z_i\sim P_0$. The expectation of the derivative matrix of the full moment

conditions at P_0 is given by

$$
\begin{aligned}
M &= E_{P_0}\left(\frac{\partial}{\partial\theta'}\mathbf{m}\left(Z_i,\theta_0\right)\right) \\
&= \begin{pmatrix} -E_{P_0}\left(hh'\right) & \mathbf{0} & \mathbf{0} \\ E_{P_0}\left(\frac{g}{1-G}\left(\mu_0\left(X\right)-\alpha\right)W'\right) & -2Q & -Q \\ \mathbf{0}' & -Q & -Q \end{pmatrix},
\end{aligned}
$$

and its inverse is

$$
M^{-1} = \begin{pmatrix} -E_{P_0}\left(hh'\right)^{-1} & \mathbf{0} & \mathbf{0} \\ -\frac{1}{Q}E_{P_0}\left(\frac{g}{1-G}\left(\mu_0\left(X\right)-\alpha\right)W'\right)E_{P_0}\left(hh'\right)^{-1} & -Q^{-1} & Q^{-1} \\ \frac{1}{Q}E_{P_0}\left(\frac{g}{1-G}\left(\mu_0\left(X\right)-\alpha\right)W'\right)E_{P_0}\left(hh'\right)^{-1} & Q^{-1} & -2Q^{-1} \end{pmatrix}.
$$

Hence,$\left(\Lambda_S M\Lambda_S'\right)^{-1}$ is obtained as

$$
\left(\Lambda_S M\Lambda_S'\right)^{-1} = \begin{pmatrix} -E_{P_0}\left(h_S h_S'\right)^{-1} & \mathbf{0} & \mathbf{0} \\ -\frac{1}{Q}E_{P_0}\left(\frac{g}{1-G}\left(\mu_0\left(X\right)-\alpha\right)W_S'\right)E_{P_0}\left(h_S h_S'\right)^{-1} & -Q^{-1} & Q^{-1} \\ \frac{1}{Q}E_{P_0}\left(\frac{g}{1-G}\left(\mu_0\left(X\right)-\alpha\right)W_S'\right)E_{P_0}\left(h_S h_S'\right)^{-1} & Q^{-1} & -2Q^{-1} \end{pmatrix}.
$$

By noting identity $E_{P_0}\left(\frac{g}{1-G}\left(\mu_0\left(X\right)-\alpha_0\right)W_S'\right) = E_{P_0}\left(\frac{D-G}{1-G}\left(\mu_0\left(X\right)-\alpha_0\right)h_S'\right)$, we can express the bottom element of $-\left(\Lambda_S M\Lambda_S'\right)^{-1}\Lambda_S\mathbf{m}\left(Z_i,\theta_0\right)$ as

$$
\begin{aligned}
&-\frac{1}{Q}E_{P_0}\left(\frac{D-G}{1-G}\left(\mu_0\left(X\right)-\alpha_0\right)h_S'\right)E_{P_0}\left(h_S h_S'\right)^{-1}h_{S,i} && \text{(A.9)} \\
&-\frac{1}{Q}\left[D_i+\left(1-D_i\right)\left(\frac{G_i}{1-G_i}\right)\right]\left(Y_i-\tau D_i-\alpha_0\right) \\
&+\frac{2}{Q}\left[D_i+\left(1-D_i\right)\left(\frac{G_i}{1-G_i}\right)\right]\left(Y_i-\tau D_i-\alpha_0\right)D_i \\
=&-\frac{1}{Q}L\left\{\left(\frac{D-G}{1-G}\right)\left[\mu_0\left(X\right)-\alpha_0\right]\middle| h_S\right\}+\frac{D_i}{Q}\left(Y_i-\mu_1\left(X\right)\right)-\frac{1-D_i}{Q}\frac{G_i}{1-G_i}\left(Y_i-\mu_0\left(X\right)\right) && \text{(A.10)} \\
&+\left(\frac{D-G}{Q}\right)\left[\mu_1\left(X\right)-\alpha_0+\frac{G}{1-G}\left(\mu_0\left(X\right)-\alpha_0\right)-\tau_0\right]+\frac{G}{Q}\left(\Delta\mu(X)-\tau_0\right).
\end{aligned}
$$

The first term of (A.10) admits the following decomposition,

$$
\begin{aligned}
&-\frac{1}{Q}L\left\{\left(\frac{D-G}{1-G}\right)\left[\mu_0\left(X\right)-\alpha_0\right]\middle| h_S\right\} \\
=&\frac{1}{Q}L\left\{(D-G)(\Delta\mu(X)-\tau_0)|h_S\right\}-\frac{1}{Q}L\left\{\left(D-G\right)\left[\mu_1\left(X\right)-\alpha_0+\frac{G}{1-G}\left(\mu_0\left(X\right)-\alpha_0\right)-\tau_0\right]\middle| h_S\right\}.
\end{aligned}
$$

Hence, we can express (A.10) as

$$
\begin{aligned}
&\frac{1}{Q}L\left\{(D-G)(\Delta\mu(X)-\tau)|h_S\right\}+\frac{1}{Q}L^{\perp}\left\{\left(D-G\right)\left[\mu_1\left(X\right)-\alpha_0+\frac{G}{1-G}\left(\mu_0\left(X\right)-\alpha_0\right)-\tau_0\right]\middle| h_S\right\} \\
&+\frac{D_i}{Q}\left(Y_i-\mu_1\left(X\right)\right)-\frac{1-D_i}{Q}\frac{G_i}{1-G_i}\left(Y_i-\mu_0\left(X\right)\right)+\frac{G}{Q}\left(\Delta\mu(X)-\tau_0\right).
\end{aligned}
$$

These five terms are mean zero and mutually uncorrelated, so the sum of their variances gives the asymptotic variance of $\sqrt{n}\left(\hat{\tau}_S - \tau_n\right)$.

Regarding the bias term, (A.8) suggests that it is given by the bottom element of the second term in the right hand side of (A.8), which is calculated as

$$\begin{aligned}
&-\frac{1}{Q}E_{P_0}\left[\frac{D-G}{1-G}\left[\mu_0\left(X\right)-\alpha_0\right]h_S'\right]E_{P_0}\left(h_S h_S'\right)^{-1}E_{P_0}\left(h_S h_{S^c}'\right)\delta_{S^c}+\frac{1}{Q}E_{P_0}\left[\frac{D-G}{1-G}\left[\mu_0\left(X\right)-\alpha_0\right]h_{S^c}'\right]\delta_{S^c}\\
=&\frac{1}{Q}E_{P_0}\left[\left\{\frac{D-G}{1-G}\left[\mu_0\left(X\right)-\alpha_0\right]-E_{P_0}\left[\frac{D-G}{1-G}\left[\mu_0\left(X\right)-\alpha_0\right]h_S'\right]E_{P_0}\left(h_S h_S'\right)^{-1}h_S\right\}h_{S^c}'\right]\delta_{S^c}\\
=&E_{P_0}\left[\frac{1}{Q}L^{\perp}\left\{\left(\frac{D-G}{1-G}\right)\left[\mu_0\left(X\right)-\alpha_0\right]\middle|h_S\right\}h_{S^c}'\right]\delta_{S^c}.
\end{aligned}$$

■

Derivation of (3.8). The asymptotic MSE of $\sqrt{n}\left(\hat{\tau}_{avg}-\tau_n\right)$ is given by the limit of

$$Var_{P_n}\left(\sqrt{n}\left(\hat{\tau}_{avg}-\tau_n\right)\right)+\left[E_{P_n}\left(\sqrt{n}\left(\hat{\tau}_{avg}-\tau_n\right)\right)\right]^2.$$

Using the formula of conditional variance, we rewrite the MSE as

$$\begin{aligned}
&E_{P_n}\left[Var_{P_n}\left(\sqrt{n}\left(\hat{\tau}_{avg}-\tau_n\right)|\hat{\delta}_{\underline{S}^c}\right)\right]+Var_{P_n}\left[E_{P_n}\left(\sqrt{n}\left(\hat{\tau}_{avg}-\tau_n\right)|\hat{\delta}_{\underline{S}^c}\right)\right]+\left[E_{P_n}\left(\sqrt{n}\left(\hat{\tau}_{avg}-\tau_n\right)\right)\right]^2\\
&=E_{P_n}\left[Var_{P_n}\left(\sqrt{n}\left(\hat{\tau}_{avg}-\tau_n\right)|\hat{\delta}_{\underline{S}^c}\right)\right]+E_{P_n}\left[E_{P_n}\left(\sqrt{n}\left(\hat{\tau}_{avg}-\tau_n\right)|\hat{\delta}_{\underline{S}^c}\right)^2\right].
\end{aligned}$$

That is, the unconditional MSE of $\sqrt{n}\left(\hat{\tau}_{avg}-\tau_n\right)$ is equal to the mean of its conditional MSE given $\hat{\delta}_{\underline{S}^c}$. From (3.7) and by noting $\sqrt{n}\left(\hat{\tau}_{avg}-\tau_n\right)=\mathbf{c}\left(\hat{\delta}_{\underline{S}^c}\right)'\hat{\mathbf{t}}_n$, the limit of the conditional MSE of $\sqrt{n}\left(\hat{\tau}_{avg,NPW}-\tau_n\right)$ given $\hat{\delta}_{\underline{S}^c}$ is written as

$$\mathbf{c}\left(\hat{\delta}_{\underline{S}^c}\right)'\left(\Omega_{22}-\Omega_{21}\Omega_{11}^{-1}\Omega_{12}\right)\mathbf{c}\left(\hat{\delta}_{\underline{S}^c}\right)+\left\{\mathbf{c}\left(\hat{\delta}_{\underline{S}^c}\right)'\left[B\delta_{\underline{S}^c}+\Omega_{21}\Omega_{11}^{-1}\left(\hat{\delta}_{\underline{S}^c}-\delta_{\underline{S}^c}\right)\right]\right\}^2.$$

Hence, it follows the quadratic form expression in the expectation in (3.8) with the corresponding weighting matrix $K\left(\hat{\delta}_{\underline{S}^c},\delta_{\underline{S}^c}\right)$. ■

Proof of Proposition 3.1. Solving the Bayes decision problem (3.9) is equivalent to solving the posterior Bayes action for every possible realization of $\hat{\delta}_{\underline{S}^c}$. Hence, we let $\hat{\delta}_{\underline{S}^c}$ be given by data, and consider minimizing the posterior risk for $c\left(\hat{\delta}_{\underline{S}^c}\right)$ subject to the normalization constraint,

$$\begin{aligned}
&\min_{\mathbf{c}\left(\hat{\delta}_{\underline{S}^c}\right)}\mathbf{c}\left(\hat{\delta}_{\underline{S}^c}\right)'E_{\delta_{\underline{S}^c}|\hat{\delta}_{\underline{S}^c}}\left[K\left(\hat{\delta}_{\underline{S}^c},\delta_{\underline{S}^c}\right)\right]\mathbf{c}\left(\hat{\delta}_{\underline{S}^c}\right),\\
&\text{s.t.}\quad \mathbf{c}\left(\hat{\delta}_{\underline{S}^c}\right)'\mathbf{1}=1,
\end{aligned}$$

If $K_{post}\left(\hat{\delta}\right)=E_{\hat{\delta}_{\underline{S}^c}|\delta_{\underline{S}^c}}\left[K\left(\hat{\delta}_{\underline{S}^c},\delta_{\underline{S}^c}\right)\right]$ is nonsingular, this constrained optimization becomes a quadratic minimization problem with a convex objective function. So, this has a unique solution and the standard

Lagrangian optimization procedure yields $\mathbf{c}^*\left(\hat{\delta}_{\underline{S}^c}\right)$ of the proposition. Note that, with proper $\mu\left(\delta_{\underline{S}^c}\right)$, the minimized Bayes risk is bounded, because, by considering weight that always assignes 1 to the largest model, we have

$$\int E_{\hat{\delta}_{\underline{S}^c}|\delta_{\underline{S}^c}}\left[\mathbf{c}^*\left(\hat{\delta}_{\underline{S}^c}\right)' K\left(\hat{\delta}_{\underline{S}^c},\delta_{\underline{S}^c}\right)\mathbf{c}^*\left(\hat{\delta}_{\underline{S}^c}\right)\right] d\mu\left(\delta_{\underline{S}^c}\right) \leqq \omega^2_{largest}\int d\mu\left(\delta_{\underline{S}^c}\right) = \omega^2_{largest} < \infty,$$

where $\omega^2_{largest}$ is the asymptotic variance of the NPW estimator in the largest model. ■

B Derivation of Equation (4.2)

We apply the formula of Proposition 2.1. Under the presented setup of Section 4.1, note that h_S of the small model is $\frac{(D-G)}{G(1-G)}$ and that of the large model is $\frac{(D-G)}{G(1-G)}(1,X)'$.

First, we compute the MSE of the small model. In the small model, straightforward calculations yield the following identities

$$\begin{aligned}
L\left\{(D-G)\left[\Delta\mu\left(X\right)-\tau_0\right]|h_{small}\right\} &= 0,\\
L^{\perp}\left\{(D-G)\left[\Delta\mu\left(X\right)-\tau_0+\frac{1-2G}{1-G}\left(\mu_0(X)-\alpha_0\right)\right]\middle| h_{small}\right\} &= \frac{(D-G)G}{(1-G)}(X-p_X)\mu,\\
L^{\perp}\left\{\frac{D-G}{1-G}\left(\mu_0(X)-\alpha_0\right)\middle| h_{small}\right\} &= -\frac{(D-G)}{(1-G)}(X-p_X)\mu
\end{aligned}$$

Hence, by denoting the first term in the right hand side of (2.4) by ω^2, the variance of the NPW estimator in the small model is given by

$$\begin{aligned}
\omega^2_{small} &= \omega^2 + \frac{1}{G^2}E_{P_0}\left[\left(\frac{(D-G)G}{(1-G)}(X-p_X)\mu\right)^2\right]\\
&= \omega^2 + \frac{G}{1-G}p_X(1-p_X)\mu^2.
\end{aligned}$$

The bias term in the small model is

$$\begin{aligned}
bias_{small} &= -\frac{1}{G}E_{P_0}\left[\left(\frac{(D-G)^2}{G(1-G)^2}X(X-p_X)\mu\right)\right]\delta\\
&= -\frac{p_X(1-p_X)}{G(1-G)}\mu\delta.
\end{aligned}$$

Thus, we obtain the MSE of the small model as

$$MSE_{small} = \omega^2 + \frac{G}{1-G}p_X(1-p_X)\mu^2 + \left[\frac{p_X(1-p_X)}{G(1-G)}\mu\delta\right]^2.$$

Next, consider the large model. We have

$$\begin{aligned}
L\left\{(D-G)\left[\Delta\mu\left(X\right)-\tau_0\right]|h_{large}\right\} &= (D-G)(X-p_X)\mu,\\
L^{\perp}\left\{(D-G)\left[\Delta\mu\left(X\right)-\tau_0+\frac{1-2G}{1-G}\left(\mu_0(X)-\alpha_0\right)\right]\middle| h_{large}\right\} &= 0,
\end{aligned}$$

where the second equality follows since $(D-G)\left[\Delta\mu(X)-\tau_0+\frac{1-2G}{1-G}(\mu_0(X)-\alpha_0)\right]$ is linear in $(D-G)$ and $(D-G)X$. The second equality and the zero bias in the large model imply that the MSE of the NPW estimator in the large model equals to (2.4),

$$\begin{aligned} MSE_{large} &= \omega^2 + \frac{1}{G^2}E_{P_0}\left[(D-G)^2(X-p_X)^2\mu^2\right] \\ &= \frac{1-G}{G}p_X(1-p_X)\mu^2. \end{aligned}$$

Hence, $MSE_{small} - MSE_{large}$ as presented in (4.2) follows.

References

[1] Abadie, A. and G.W. Imbens (2006), "Large Sample Properties of Matching Estimators for Average Treatment Effects," *Econometrica*, 74, 235-267.

[2] Belloni, A., V. Chernozhukov, and C. Hansen (2013), "Inference on Treatment Effects After Selection Amongst High-dimensional Controls," *Review of Economic Studies*, forthcoming.

[3] Busso, M., J. DiNardo, and J. McCrary (2011), "New Evidence on the Finite Sample Properties of Propensity Score Reweighting and Matching Estimators," *IZA Discussion Paper*, No. 3998.

[4] Chen, X., H. Hong, and A. Tarozzi (2008), "Semiparametric Efficiency in GMM Models with Auxiliary Data," *Annals of Statistics*, 36, 808-843.

[5] Claeskens, G. and R.J. Carroll (2007), "An Asymptotic Theory for Model Selection Inference in General Semiparametric Problems," *Biometrika*, 94, 249-265.

[6] Claeskens, G. and N.L. Hjort (2003), "The Focussed Information Criterion," *Journal of the American Statistical Association*, 98, 900-916 (with discussion).

[7] Crump, R.K., J. Hotz, G.W. Imbens, and O.A. Mitnik (2009), "Dealing with Limited Overlap in Estimation of Average Treatment Effects," *Biometrika*, 96, 187-199.

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

[9] DiTraglia, F. J. (2013), "Using Invalid Instruments on Purpose: Focused Moment Selection and Averaging for GMM," *unpublished manuscript*, University of Pennsylvania.

[10] Graham, B.S., C.C. de Xavier Pinto, and D. Egel (2011), "Efficient Estimation of Data Combination Models by the Method of Auxiliary-to-study Tilting (AST)," *NBER Working Paper*, No. 16928.

[11] Graham, B.S., C.C. de Xavier Pinto, and D. Egel (2012), "Inverse Probability Tilting for Moment Condition Models with Missing Data," *Review of Economic Studies*, 79, 1053-1079.

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

[13] Hahn, J. (2004), "Functional Restriction and Efficiency in Causal Inference," *Review of Economics and Statistics*, 86, 1, 73-76.

[14] Hansen, B.E. (2005), "Challenges for Econometric Model Selection," *Econometric Theory*, 21, 60-68.

[15] Hansen, B.E. (2007), "Least Squares Model Averaging," *Econometrica*, 75, 1175-1189.

[16] Hansen, B.E., and J.S. Racine (2012), "Jackknife Model Averaging," *Journal of Econometrics*, 167, 38–46.

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

[18] 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.

[19] Hirano, K. and J.R. Porter (2009), "Asymptotics for Statistical Treatment Rules," *Econometrica*, 77, 1683-1701.

[20] Hjort, N.L. and G. Claeskens (2003), "Frequentist Model Average Estimators," *Journal of the American Statistical Association*, 98, 879-899 (with discussion).

[21] Hjort, N.L. and G. Claeskens (2006), "Focussed Information Criteria and Model Averaging for Cox's Hazard Regression Model," *Journal of the American Statistical Association*, 101, 1449-1464.

[22] Hjort, N.L. and G. Claeskens (2008), *Model Selection and Model Averaging*, Cambridge University Press, Cambridge, UK.

[23] Ichimura, H. and O. Linton (2001), "Asymptotic Expansions for Some Semiparametric Program Evaluation Estimators," *Cemmap working paper*, 01/04.

[24] Imbens, G.W. (2004), "Nonparametric Estimation of Average Treatment Effects Under Exogeneity: A Review," *Review of Economics and Statistics*, 86, 1, 4-29.

[25] Imbens, G.W., W. Newey, and G. Ridder (2005), "Mean-square-error Calculations for Average Treatment Effects," *IEPR Working Paper* 05.34, University of Southern California.

[26] Khan, S. and E. Tamer (2010), "Irregular Identification, Support Conditions, and Inverse Weight Estimation," *Econometrica*, Vol 78, 6, 2021-2042.

[27] LaLonde, R.J. (1986), "Evaluating the Econometric Evaluation of Training Programs with Experimental Data," *American Economic Review*, Vol. 76, 4, 604-620.

[28] Lu, Xun. (2013), "A Covariate Selection Criterion for Estimation of Treatment Effects," *unpublished manuscript*, Hong Kong University of Science and Technology.

[29] Magnus, J.R., O. Powell, and P. Prüfer (2010), "A Comparison of Two Model Averaging Techniques with an Application to Growth Empirics," *Journal of Econometrics*, 154, 139-153.

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

[31] Song, K. (2013), "Point Decisions for Interval Identified Parameters," *Econometric Theory*, forthcoming.

[32] Sueishi, N. (2013), "Empirical Likelihood-Based Focused Information Criterion and Model Averaging," *unpublished manuscript*, Kyoto University.

[33] Tibshirani, R. (1996), "Regression Shrinkage and Selection via the Lasso," *Journal of the Royal Statistical Society, Series B*, 58, 267-288.

[34] van der Vaart, A. (1998), *Asymptotic Statistics*, Cambridge University Press, New York.

[35] van der Vaart, A. W., and J.A. Wellner (1996): *Weak Convergence and Empirical Processes: With Applications to Statistics,* New York: Springer.

[36] Vansteelandt, S., M. Bekaert, and G. Claeskens (2012), "On Model Selection and Model Misspecification in Causal Inference," *Statistical Methods in Medical Research*, 21, 7-30.

[37] Wan, A.T.K., X. Zhang, and G. Zou (2010), "Least Squares Model Averaging by Mallows Criterion," *Journal of Econometrics*, 156, 277–283.

[38] Wooldridge, J.M. (2002), *Econometric Analysis of Cross Section and Panel Data*, MIT Press, Cambridge, MA.

[39] Zhang, X. and H. Liang (2011), "Focused Information Criterion and Model Averaging for Generalized Additive Partial Linear Models," *Annals of Statistics*, 39, 174-200.