cemmap
centre for microdata methods and practice

Model Averaging in Semiparametric Estimation of Treatment Effects

Toru Kitagawa
Chris Muris

The Institute for Fiscal Studies
Department of Economics, UCL

cemmap working paper CWP46/15

E·S·R·C ECONOMIC & SOCIAL RESEARCH COUNCIL

An ESRC Research Centre

Model Averaging in Semiparametric Estimation of Treatment Effects*

Toru Kitagawa† and Chris Muris‡

August 11, 2015

Abstract

In the practice of program evaluation, choosing the covariates and the functional form of the propensity score is an important choice that the researchers make when estimating treatment effects. This paper proposes a data-driven way of averaging the estimators over the candidate specifications in order to resolve the issue of specification uncertainty in the propensity score weighting estimation of the average treatment effects for treated (ATT). The proposed 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 localization parameters. Analytical comparisons of the Bayes asymptotic MSE show that the averaging estimator outperforms post model selection estimators and the estimators in any of the candidate models. Our Monte Carlo studies confirm these theoretical results and illustrate the size of the MSE gains from averaging. We apply the averaging procedure to evaluate the effect of the labor market program analyzed in LaLonde (1986).

Keywords: Treatment effects, Propensity score, Model averaging, Limit experiment.
JEL Classification: C13, C21, C52.

*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 thank an associate editor and three referees for their thorough reading and constructive comments on an earlier draft. 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 the estimation of the causal impact of various social programs. When the exposure to or participation in 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 modeling 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.

In order to solve this practical issue of specification uncertainty in causal inference, this paper proposes a method to construct a best causal effect estimator by averaging the estimators obtained in different candidate specifications. We focus on the average treatment effect for the treated (ATT) as the estimand of interest, and consider the averaging the propensity score weighting estimators. Building on the idea of frequentist model averaging proposed by Hjort and Claeskens (2003) and Hansen (2007), our model averaging procedure aims to construct a point estimator for ATT in the form of a weighted average of the ATT estimators in the candidate models, where the weights are optimally chosen in a data-driven way to minimize the mean squared error (MSE) of the averaged estimator.

The model averaging procedure proposed in this paper proceeds 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,

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

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; $W\left(X\right)$ is a vector of functions of the pre-treatment covariates X that can contain interactions and nonlinear transformations of X; and $G\left(\cdot\right)$ is a known link function such as the logit function. Candidate models to be averaged 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. This local misspecification framework leads to a useful approximation of the MSE of an averaging estimator as a function of δ. Since δ remains unknown even in large samples, the optimal averaging weights depend crucially on how the non-vanishing uncertainty about the localization parameters is dealt with. 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 a 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), in which the treatment of the localization parameters, to the best of our knowledge, lacks a decision-theoretic optimality argument.

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 the 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 there is evidence from simulation studies that suggests that the finite sample performance of the NPW estimator is excellent (see Busso, DiNardo, and McCrary (2014)). The main reason that we focus on the ATT rather than the average treatment effect for the whole population (ATE) closely relates to the fact that the semiparametric efficiency bound for the 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 the ATT by having a 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 trade-off in the ATT estimation,[1] which the averaging weights aim to optimally balance out.

[1]This bias-variance trade-off is not available in the propensity score weighted estimation for ATE as shown in Section 2 below.

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 NPW estimators in any of the candidate models including the MSE minimizing one. In our Monte Carlo specifications, this MSE gain from averaging relative to a correctly specified largest model is about 10% for a large range of localization parameter values. To illustrate the use of our model averaging procedure, we apply it to the data set used by LaLonde (1986) to evaluate a job-training program in the United States.

1.1 Related Literature

The averaging procedure proposed in this paper contributes to the growing literature of frequentist model averaging. The frequentist model averaging that targets to minimize the MSE for a parameter of interest is pursued by Hjort and Claeskens (2003) in general parametric models. This paper extends their model averaging framework to the context of semiparametric estimation of causal effect. In the least squares regression context, frequentist model averaging with the MSE criterion of the entire regression function (integrated MSE) is analyzed by Hansen (2007, 2014a), Wan, Zhang, and Zou (2010), and Hansen and Racine (2012), Liu and Okui (forthcoming), among others. Magnus, Powell, and Prüfer (2010) propose 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 focused information criterion (FIC)-based model averaging for generalized method of moment estimators, with primary applications to linear instrumental variable models. Liu (2013) proposes a novel procedure for conducting inference for FIC in linear models. Lu (2013) considers averaging semiparametric estimators for the 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. Since averaging results in shrinking the estimator in the largest model toward the estimators in smaller models, the averaging estimator can be interpreted as a shrinkage estimator, which has a long history in statistics since James and Stein (1961). Using a local asymptotic framework in a general parametric model, Hansen (2014b) proposes a shrinkage estimator that uniformly dominates the maximum likelihood estimator in the largest model. Cheng, Liao, and Shi (2015) show the uniform dominance property of the shrinkage estimator in the context of generalized

method of moments. In contrast to the shrinkage analysis that generally focuses on estimation of multi-dimensional parameters, the parameter of interest in the current context is one-dimensional.

Model averaging can be seen as a generalization of model selection, since the latter restricts the averaging weights to ones and zeros. In this regard, the MSE performance of the averaging procedure outperforms any of the model selection procedure that relies on the same MSE criterion as our procedure, e.g., model selection based on FIC proposed by Claeskens and Hjort (2003) in parametric models. FIC-based model selection in semiparametric models are considered in Hjort and Claeskens (2006), Claeskens and Carroll (2007), and Zhang and Liang (2011), among others. Vansteelandt, Bekaert, and Claeskens (2012) propose a FIC-based variable selection procedure for the average treatment effect as a focused parameter in a parametric context. Millimet and Tchernis (2009) provide some simulation evidence in favor of selecting parsimonious models. When the propensity scores and/or the outcome regression equations are nonparametrically estimated, the problem of specification choice is reduced to the problem of selecting 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 explicitly aiming to minimize the MSE of the 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. In practical terms, our parametric restriction is convenient to deal with multidimensional covariates. Also, the proposed procedure does not require a preliminary nonparametric estimate of unknown functions (cf. Ichimura and Linton (2001)). Our approach, however, 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 model averaging approach to covariate selection differs from their Lasso approach in terms of the scope of applications and the notion of optimality that these procedures aim to achieve asymptotically. First, our averaging procedure mainly concerns 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 situations where the number of regressors is equal to even larger than the sample size. Second, optimality of our 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. In addition, 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 ATE and ATT. See also Farrell (2013), who derives uniformly valid inference procedures in a similar setup.

Our derivation of the optimal averaging weights solves a Bayes optimal statistical decision in a limit normal experiment, which is different from Hjort and Claeskens's proposal to base the weights on plug-in estimators. In econometrics, decision-theoretic analyses in limit experiments have been conducted in various contexts; see Hirano and Porter (2009) for the treatment choice problem, and Song (2014) for the point estimation problem for interval-identified parameters.

1.2 Plan of the Paper

In Section 2, we introduce the local misspecification framework for ATE and ATT estimation, and derive the asymptotic MSEs for the NPW estimators of the candidate models. We also examine the bias-variance trade-off between large and parsimonious models through the analytical expression of the asymptotic MSEs. In Section 3, we propose our optimal averaging procedure that minimizes the Bayes risk (a weighted average of MSE) criterion in the limit experiment. The results of our Monte Carlo studies are provided in Section 4. Section 5 applies our averaging procedure to LaLonde's (1986) data set on the National Supported Work Demonstration job-training program. Section 6 concludes. All proofs of the propositions and auxiliary lemmas are collected in Appendix A.

2 Estimation of Causal Effects with Locally Misspecified Propensity Scores

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 $Y_i \in \mathbb{R}$, 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 satisfies $Y_i = D_i Y_i(1) + (1 - D_i) Y_i(0)$. The population average treatment effect (ATE) and the average treatment effect for treated (ATT) when $(Y(1), Y(0), D, X) \sim P$ are denoted by $\tau^{ATE} = E_P(Y(1) - Y(0))$ and $\tau^{ATT} = E_P(Y(1) - Y(0)|D = 1)$, respectively.

The starting point of our averaging procedure is to specify a most complicated specification for the propensity score function, which we refer to as the *largest model*. Let $W(X) \in \mathbb{R}^K$ 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 will use a short-hand notation, $W_i = W(X_i)$, as far as no confusion arises.

Each candidate specification for the propensity score corresponds to a subvector of $W(X)$ used in the propensity score estimation. We index by S a selection of covariates of W. The number of covariates included in specification S is denoted by $|S|$. We denote the set of candidate specifications by $\mathcal{M}$ and the number of models in it by $|\mathcal{M}|$. The set $\mathcal{M}$ does not have to exhaust all the possible subset vectors of $W(X)$. For example, some regressors can be included in all the specifications if they are believed to be important in predicting treatment status. Let $\underline{S} = \cap\{S : S \in \mathcal{M}\}$ be the set of covariates that appear in every candidate model. 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, $\underline{S}^c$ is the set of covariates that are excluded in some candidate model.

The next set of assumptions characterizes sequences of data generating processes $\{P_{n,\delta} : n = 1, 2, \dots\}$. It will form the basis for our local asymptotic analysis, and for the limiting experiment that gives rise to our optimal averaging procedure.

Assumption DGP:

(i) *(Unconfoundedness) The joint distribution of* $(Y(1), Y(0), D, X)$ *satisfies*

$$P_{n,\delta}(Y(1), Y(0), D, X) = P_0(Y(1), Y(0)|X) \cdot P_{n,\delta}(D|X) \cdot P_0(X), \tag{2.1}$$

where $P_0(Y(1), Y(0)|X)$ *is the conditional distribution of potential outcomes given the full set of covariates* X *and* $P_0(X)$ *is the marginal distribution of* X*, which are independent of* n*.*

(ii) *(Propensity score specification)* $P_{n,\delta}(D|X)$ *depends on the sample size and* $P_{n,\delta}(D = 1|X = x) = G\left(W(x)'\gamma_n\right)$*,* $\gamma_n \in \mathbb{R}^K$ *with a known monotone and twice continuously differentiable link function* $G(\cdot)$*.*

(iii) *(Localized parameter sequence)* $\gamma_n = \gamma_0 + n^{-1/2}\delta$*, where* $\gamma_0 \in \mathbb{R}^K$ *is a benchmark centering value of the coefficient vector and* $\delta \in \mathbb{R}^K$ *is the localization parameter.*

(iv) *(Local misspecification) Entries of* γ_0 *are zero if the corresponding regressors in* W *are excluded in some candidate specification in* $\mathcal{M}$*.*

Following Claeskens and Hjort (2003), we specify the data generating processes to be drifting with n. Note that the only drifting component is the propensity score and the other parts of the data

generating process do not change with n.[2] Decomposition (2.1) assumes *unconfoundedness* (selection on observables) of the treatment assignment with the *full* set of covariates, i.e., $(Y(1), Y(0))$ is statistically independent of D conditional on X.

Assumption DGP (ii) states that the propensity score has a parametric single index specification with a known link function. The literature on semiparametric estimation of average causal effects 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, what Assumption DGP (ii) essentially excludes are cases with a number of series terms comparable with the sample size.

Assumption DGP (iii) introduces a drifting sequence of parameters with localization parameters δ. Assumption DGP (iv) implies that the largest true model shrinks to the parsimonious submodels where only a subset of $W(X)$ is used in the propensity score estimation. In this sense, the smaller models are locally misspecified and the value of localization parameters δ measures the degree of misspecification in terms of the coefficient values. The joint distribution of $(Y(1), Y(0), D, X)$ when γ is set at γ_0 (i.e., $\delta = 0$) is denoted by P_0.

Assumption DGP (i) implies that the ATE parameter does not depend on n, $\tau_0^{ATE} \equiv E_{P_0}(Y(1) - Y(0))$, whereas the ATT parameter does, $\tau_n^{ATT} \equiv E_{P_n}(Y(1) - Y(0)|D = 1)$, since the marginal distribution of D depends on γ_n. Under unconfoundedness, the ATE and ATT parameters satisfy the following moment conditions: at every n,

$$E_{P_{n,\delta}}\left[\frac{D_i Y_i}{G\left(W(x)'\gamma_n\right)} - \frac{(1-D_i)Y_i}{1-G(W_i'\gamma_n)} - \tau_0^{ATE}\right] = 0,$$

$$E_{P_{n,\delta}}\left[\frac{D_i Y_i}{Q_n} - \frac{G(W_i'\gamma_n)(1-D_i)Y_i}{Q_n(1-G(W_i'\gamma_n))} - \tau_n^{ATT}\right] = 0.$$

where $E_{P_{n,\delta}}$ is the expectation with respect to the data generating process $P_{n,\delta}$ defined in (2.1) and $Q_n \equiv P_{n,\delta}(D = 1)$.

Let $\hat{\gamma}$ be the maximum likelihood estimator for γ_n obtained from the parametric binary regression based on Assumption DGP (ii), and $\hat{Q} = \frac{1}{n}\sum_{i=1}^n D_i$. The *normalized propensity score weight (NPW) estimators* for the ATE and ATT in the largest model are

$$\hat{\tau}^{ATE} = \sum_{i=1}^n \left(\frac{D_i}{G(W_i'\hat{\gamma})}Y_i \Big/ \sum_{i=1}^n \frac{D_i}{G(W_i'\hat{\gamma})} - \frac{(1-D_i)}{(1-G(W_i'\hat{\gamma}))}Y_i \Big/ \sum_{i=1}^n \frac{(1-D_i)}{(1-G(W_i'\hat{\gamma}))}\right), \quad (2.2)$$

[2]We can allow the potential outcome distribution and the marginal distribution of X to drift with the sample size without affecting the analytical results and the model selection/averaging procedures in this paper. However, for the sake of parsimony of the exposition, we will leave them independent of n in what follows.

$$\hat{\tau}^{ATT} = \sum_{i=1}^{n} \left(\frac{D_i Y_i}{\sum_{i=1}^{n} D_i} - \frac{G\left(W_i'\hat{\gamma}\right)\left(1-D_i\right)}{\left(1-G\left(W_i'\hat{\gamma}\right)\right)} Y_i \Bigg/ \sum_{i=1}^{n} \frac{G\left(W_i'\hat{\gamma}\right)\left(1-D_i\right)}{\left(1-G\left(W_i'\hat{\gamma}\right)\right)} \right), \tag{2.3}$$

where the summation terms in the denominators guarantee that the weights that multiply the observed outcomes sum up to one.

The $|S| \times 1$ subvectors of W and γ corresponding to the selected covariates in model S are denoted by W_S and γ_S, respectively. We define the $|S| \times K$ matrix π_S such that pre-multiplying a $K \times 1$ vector by π_S yields the subvector corresponding to selection S, i.e., $\pi_S W = W_S$ and $\pi_S \gamma = \gamma_S$ hold. Given a selection of covariates S, let $\hat{\tau}_S^{ATE}$ and $\hat{\tau}_S^{ATT}$ be the NPW-ATE and NPW-ATT estimators when W_S is included in the estimation of the parametric propensity score, i.e.,

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

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

where $\hat{\gamma}_S$ is the maximum likelihood estimator for γ_S obtained in the first stage propensity score regression of D_i on $W_{S,i}$.[3]

In addition to Assumption DGP, we impose the following regularity conditions on the sequence of DGPs to ensure $\sqrt{n}-$local asymptotic normality of the estimators:

Assumption REG: *(Regularity conditions and overlap) Let $\Gamma \subset \mathbb{R}^K$ be the parameter space for γ.*

(i) Γ is compact and γ_0 is in the interior of Γ.

(ii) Let $l(Z,\gamma)$ denote the one-observation log likelihood for γ in the first stage propensity score estimation, where $Z = (Y, D, W(X))$. The largest model and the candidate submodels are globally identified in the sense that, for every $\epsilon > 0$, there exists constant $\lambda_\epsilon > 0$ such that

$$E_{P_{n,\delta}}\left[l(Z,\gamma_n)\right] > \sup_{\gamma \in \Gamma : \|\gamma - \gamma_n\| > \epsilon} E_{P_{n,\delta}}\left[l(Z,\gamma)\right] + \lambda_\epsilon$$

[3] As an alternative to the NPW estimator in model S, we may consider an overidentified GMM estimator. For instance, using the moment conditions $\mathbf{m}_i^{ATT}(\theta)$ to be defined in Section 3 and an optimal choice of weighting matrix Σ, a GMM estimator for τ^{ATT} in model S minimizes $\left(\frac{1}{n}\sum \mathbf{m}_i^{ATT}(\theta)\right)' \Sigma^{-1} \left(\frac{1}{n}\sum \mathbf{m}_i^{ATT}(\theta)\right)$ subject to $\gamma_{S^c} = 0$. Although this GMM estimator leads to improvement of asymptotic variance, its computation is not as simple as the NPW estimator considered here. We therefore do not consider such overidentified GMM estimators in our analysis.

and

$$E_{P_{n,\delta}}\left[l(Z,\tilde{\gamma}_n^S)\right] > \sup_{\gamma\in\Gamma_S:\|\gamma-\tilde{\gamma}_n^S\|>\epsilon} E_{P_{n,\delta}}\left[l(Z,\gamma)\right] + \lambda_\epsilon$$

hold for all n and $S \in \mathcal{M}$, where Γ_S is the constrained parameter space for γ in model S, $\Gamma_S = \{\gamma \in \Gamma : \gamma_{S^c} = \mathbf{0}\}$, and $\tilde{\gamma}_{n,S}$ is the pseudo-true value in model S defined by $\tilde{\gamma}_{n,S} = \arg\max_{\gamma\in\Gamma_S} E_{P_{n,\delta}}\left[l(Z,\gamma)\right]$. The limiting information matrix for γ,

$$\mathcal{I}_\gamma \equiv E_{P_0}\left[\frac{g(W'\gamma_0)}{G(W'\gamma_0)(1-G(W'\gamma_0))}WW'\right]$$

is bounded and nonsingular.

(iii) Let $g(a) \equiv \frac{d}{da}G(a)$ and denote the Euclidean metric of W by $\|W\|$. $E_{P_0}\left[\sup_{\gamma\in\Gamma}\frac{g(W'\gamma)}{G(W'\gamma)}\|W\|\right] < \infty$ and $E_{P_0}\left[\sup_{\gamma\in\Gamma}\frac{g(W'\gamma)}{1-G(W'\gamma)}\|W\|\right] < \infty$.

(iv) Let W_k, $k \in \{1,\ldots,K\}$ be the k-th element of W and$[WW']_{kl}$, $k,l \in \{1,\ldots,K\}$, be the (k,l)-element of matrix $W(X)W(X)'$. There exist open neighborhood $\mathcal{N}$ of γ_0 and $\lambda > 0$ such that

$$E_{P_0}\left[\sup_{\gamma\in\mathcal{N}}\left|\frac{Y_1}{G(W'\gamma)}\right|^{2+\lambda}\right] < \infty,\ E_{P_0}\left[\sup_{\gamma\in\mathcal{N}}\left|\frac{Y_0}{1-G(W'\gamma)}\right|^{2+\lambda}\right] < \infty,$$
$$E_{P_0}\left[\sup_{\gamma\in\mathcal{N}}\left|\frac{Y_1W_k}{G(W'\gamma)^2}\right|^{1+\lambda}\right] < \infty,\ E_{P_0}\left[\sup_{\gamma\in\mathcal{N}}\left|\frac{Y_0W_k}{[1-G(W'\gamma)]^2}\right|^{1+\lambda}\right] < \infty,$$
$$E_{P_0}\left[\sup_{\gamma\in\mathcal{N}}\left|\frac{[WW']_{kl}}{[G(W'\gamma)(1-G(W'\gamma))]^2}\right|^{1+\lambda}\right] < \infty, \quad \text{for all } k,l \in \{1,\ldots,K\}.$$

Assumption REG (iii) and (iv) imply the overlap condition, $0 < G(W(x)'\gamma) < 1$ for almost every $x \in \mathcal{X}$, which is necessary for identification of ATE. The $\sqrt{n}$-asymptotic normality requires the additional conditions that restrict the tails of the marginal distribution of W and the distribution of the propensity scores near zero and one in case $G(\cdot)$ asymptotes to zero and one. Imposing these overlap conditions 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.)

Let $E_{P_0}(\cdot)$ and $Var_{P_0}(\cdot)$ be the expectation and variance at probability law P_0. In what follows, $T \overset{P_{n,\delta}}{\to} c$, or, equivalently, $T - c = o_{P_{n,\delta}}(1)$ means that the statistic T converges in probability to c along $\{P_{n,\delta}\}$, i.e., $\lim_{n\to\infty} P_{n,\delta}\left(|T-c| > \epsilon\right) = 0$ for any $\epsilon > 0$. We use $T \overset{P_{n,\delta}}{\rightsquigarrow} \mathcal{N}(\mu.\Sigma)$ to mean that the statistic (vector) T converges in distribution along $\{P_{n,\delta}\}$ to a normal distribution with mean

μ and covariance matrix Σ, i.e, $P_{n,\delta}(T \leq s) \to \Phi_{\mu,\Sigma}(s)$ as $n \to \infty$ for all $s \in \mathbb{R}^{dim(T)}$, where $\Phi_{\mu,\Sigma}(\cdot)$ is the cumulative distribution function of $\mathcal{N}(\mu.\Sigma)$. In addition, the following notation is used:

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

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. The following proposition derives the asymptotic distribution of the NPW estimators for each submodel.

Proposition 2.1 *Suppose Assumptions DGP and REG. 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_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 distributions of $\hat{\tau}_S^{ATE}$ and $\hat{\tau}_S^{ATT}$ along $\{P_{n,\delta}\}$ are

$$
\sqrt{n}\left(\hat{\tau}_S^{ATE} - \tau^{ATE}\right) \overset{P_{n,\delta}}{\rightsquigarrow} \mathcal{N}\left(0, \omega_{ATE,S}^2\right) + bias_{ATE,S}(\delta)
$$

$$
\sqrt{n}\left(\hat{\tau}_S^{ATT} - \tau_n^{ATT}\right) \overset{P_{n,\delta}}{\rightsquigarrow} \mathcal{N}\left(0, \omega_{ATT,S}^2\right) + bias_{ATT,S}(\delta),
$$

where

$$
\omega_{ATE,S}^2 = SEB_{ATE} + E_{P_0}\left[L^{\perp}\left\{\frac{D-G}{G}\left(\mu_1(X) - \tau_0^{ATE} - \mu_0\right) + \frac{D-G}{1-G}\left(\mu_0(X) - \mu_0\right)\middle| h_S\right\}^2\right], \tag{2.4}
$$

$$
bias_{ATE,S}(\delta) = E_{P_0}\left[L^{\perp}\left\{\frac{D-G}{G}\left(\mu_1(X) - \tau_0^{ATE} - \mu_0\right) + \frac{D-G}{1-G}\left(\mu_0(X) - \mu_0\right)\middle| h_S\right\} h_{S^c}'\right] \delta_{S^c}. \tag{2.5}
$$

$$\omega^2_{ATT,S} = SEB_{ATT,S} + \frac{1}{Q^2} E_{P_0}\left[L^{\perp}\left\{ (D-G)\left[\Delta\mu(X) - \tau_0^{ATT} + \frac{1-2G}{1-G}\left(\mu_0(X) - \alpha_0\right)\right] \middle| h_S \right\}^2 \right], \tag{2.6}$$

$$bias_{ATT,S}(\delta) = \frac{1}{Q} E_{P_0}\left[L^{\perp}\left\{ \left(\frac{D-G}{1-G}\right)\left[\mu_0(X) - \alpha_0\right] \middle| h_S \right\} h'_{S^c} \right] \delta_{S^c}. \tag{2.7}$$

where SEB_{ATE} is the semiparametric efficiency bound for τ^{ATE} obtained by Hahn (1998),

$$SEB_{ATE} = E_{P_0}\left[\frac{\sigma_1^2(X)}{G} + \frac{\sigma_0^2(X)}{1-G} + \left(\Delta\mu(X) - \tau_0^{ATE}\right)^2 \right],$$

and $SEB_{ATT,S}$ is the semiparametric efficiency bound for τ^{ATT} obtained by Graham, de Xavier Pinto, and Egel (2012) under the a priori restriction that the propensity score is parametric and the relevant regressors are W_S, i.e., $P(D=1|X) = G(W_S'\gamma_S)$,

$$\begin{aligned} SEB_{ATT,S} = {} & E_{P_0}\left[\left(\frac{G}{Q}\right)^2 \left\{ \frac{\sigma_1^2(X)}{G} + \frac{\sigma_0^2(X)}{1-G} + \left(\Delta\mu(X) - \tau_0^{ATT}\right)^2 \right\} \right] \\ & + \frac{1}{Q^2} E_{P_0}\left[L\left\{ (D-G)\left[\Delta\mu(X) - \tau_0^{ATT}\right] | h_S \right\}^2 \right]. \end{aligned} \tag{2.8}$$

Proof. See Appendix A. ■

Before discussing the analytical insights from this proposition, it is worth clarifying the motivation of the local asymptotic analysis in the current context. The goal of our analysis is to obtain an estimator that optimally balances out the finite sample bias-variance trade-off across small to large models. For this purpose, a sequence of DGPs (specified in Assumption DGP) is used as a device for deriving a class of δ-indexed sampling distributions of the NPW estimators, in which the variance and bias approximations of the estimators appear at the same stochastic order.[4] Since consistent estimation of δ is not feasible, a value of δ that gives accurate MSE approximation in a given situation remains unknown even in large n. Accordingly, unless one model dominates the others uniformly over δ, a data-driven way of averaging the models involves a non-trivial step of handling the uncertainty of δ. We discuss this in detail in Section 3.

The following remarks summarize some useful analytical insights about the bias-variance trade-offs in the NPW estimators.

Remark 2.1 *The variance of the submodel NPW-ATE estimator (2.4) consists of the semiparametric efficiency bound for ATE derived by Hahn (1998), which does not depend on S, and the*

[4] If we consider a type of asymptotics where n increases to infinity with a fixed DGP, we would obtain a nonzero bias of a submodel estimator $\hat{\tau}_S$ that always has a larger stochastic order than the variance irrespective of the size of misspecification. Such asymptotics may provide a poor approximation for the finite sample MSEs for submodels that are only slightly misspecified.

variance of the residuals from a certain linear projection 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 is equal to the dimension of W_S implies that the variance of the residuals is monotonically decreasing in S, implying that the asymptotic variance of $\hat{\tau}_S^{ATE}$ monotonically decreases as more regressors are included. The bias term in (2.5) is zero in the largest model. Therefore, for every δ including $\delta = 0$, the largest model is optimal in terms of the asymptotic MSE. This somewhat counter-intuitive result is in line with the well-known "propensity score paradox"[5] discussed in e.g. Hirano, Imbens, and Ridder (2003), Graham, de Xavier Pinto, and D. Egel (2012).

Remark 2.2 *In contrast to the asymptotic variance for the ATE, the asymptotic variance of the submodel NPW-ATT estimator (2.6) is non-monotonic in S. Since $SEB_{ATT,S}$ depends on S through the variance of the linear projection of $(D-G)\left[\Delta\mu(X) - \tau_0^{ATT}\right]$ onto h_S, $SEB_{ATT,S}$ weakly monotonically increases as more regressors are included in the propensity score, i.e., $SEB_{ATT,S} \leq SEB_{ATT,S'}$ whenever $S \subset S'$. As in the ATE case, the second term of (2.6), which captures the inefficiency of the NPW-ATT estimators relative to the semiparametric variance bound, monotonically decreases with the dimension of W_S whenever $S \subset S'$. As a whole, whether including more regressors in the propensity score inflates the variance of $\hat{\tau}_S^{ATT}$ depends on which of the two effects (inflation of $SEB_{ATT,S}$ versus the reduction of relative inefficiency) dominates.*[6]

As in the ATE case, the bias term shown in (2.7) is given by an inner product of δ_{S^c} and the correlation vector of h_{S^c} with a certain linear projection residual. 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 for predicting treatment status. Thus, there is no general monotonic relationship available between the squared bias and the number of included regressors.

Remark 2.3 *As shown by the relative inefficiency terms in (2.4) and (2.6), the NPW estimators are not semiparametrically efficient even when the propensity score specification in the submodel*

[5]The propensity score paradox states that even when the knowledge of propensity score specification is available, using estimated propensity scores leads to a smaller asymptotic variance of the propensity score weighted ATE estimator. In the context of variable selection, this means even though some covariates do not appear in the true propensity score, including them in the propensity score estimation improves the variance of the subsequent propensity score weighted ATE estimator as far as they help to predict the potential outcomes.

[6]In the special case where the treatment effects are homogeneous, i.e., $\Delta\mu(X) = \tau_0^{ATT}$ for all X, the first component in the variance expression $SEB_{ATT,S}$ no longer depends on S, so that adding more regressors never inflates the variance of the NPW-ATT 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.

is correct. Estimation methods that lead to semiparametrically efficient ATE and ATT estimators with the finite number of moment conditions are known in the literature. For instance, Graham et al. (2011) propose the Auxiliary-to-Study Tilting (AST) estimator for the ATT that can achieve $SEB_{ATT,S}$ 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 averaging for the AST estimators can be developed along the same line of analysis given in the next section.

3 Frequentist Model Averaging for ATT Estimation

As discussed in Remark 2.2, the presence of treatment effect heterogeneity (i.e., $\Delta\mu(X)$ is not a constant) lead to nontrivial variance-bias trade offs between the small and large models when we approximate the MSEs of the NPW-ATT estimators using a local asymptotic framework. As a result, an optimal selection of regressors that minimizes the MSE of $\hat{\tau}_S^{ATT}$ can be a proper subset of the regressors in the largest model. In contrast, such a bias-variance trade-off does not arise for the ATE-NPW estimator (see Remark 2.1). For this reason, our development of model averaging procedure focuses exclusively on the ATT.

Consider an estimator for the ATT of the following averaging form,

$$\hat{\tau}_{avg}^{ATT} = \sum_{S \in \mathcal{M}} \hat{c}_S \hat{\tau}_S^{ATT}, \tag{3.1}$$

where $\hat{\mathbf{c}} \equiv (\hat{c}_S : S \in \mathcal{M})$ is an $|\mathcal{M}| \times 1$ vector of data-dependent weights assigned to each candidate model which satisfies $\sum_{S \in \mathcal{M}} \hat{c}_S = 1$.[7] By allowing some $\hat{c}_S$ to be negative. we obtain optimal weights as an interior solution with a closed-form expression, and we can potentially lower the asymptotic MSE of $\hat{\tau}_{avg}^{ATT}$ compared to the case where the weights are constrained to be non-negative.

3.1 Bayes Asymptotic Risk and Optimal Averaging

To facilitate the presentation, we formulate the NPW-ATT estimation by the following set of moment conditions (see also Busso et al. (2014)):

$$E_{P_{n,\delta}}\left[\mathbf{m}_i^{ATT}(\theta_n)\right] = \mathbf{0},$$

[7] As an alternative class of averaging estimators, one could consider the NPW-ATT estimator with averaged propensity scores plugged in. Analyzing optimal averaging weights for this class of estimators is beyond the scope of this paper.

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

where $\theta_n = \left(\gamma_n', \alpha_n, \tau_n^{ATT}\right)'$ and $\alpha_n = E_{P_{n,\delta}}(Y(0)|D=1)$. Let

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

which, under Assumptions DGP and REG, we can consistently estimate by

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

where $\hat{\theta} = \left(\hat{\gamma}', \hat{\alpha}, \hat{\tau}^{ATT}\right)'$ is the estimator for θ in the largest model (Lemma A.2 in Appendix A). 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 $\sqrt{n}\left(\hat{\tau}_S^{ATT} - \tau_n^{ATT}\right)$ can be written as

$$\omega_{ATT,S}^2 = \text{the final element in the bottom row of} \quad (3.2)$$
$$\left(\Lambda_S M^{ATT} \Lambda_S'\right)^{-1} \Lambda_S \Sigma^{ATT} \Lambda_S' \left(\Lambda_S (M^{ATT})' \Lambda_S'\right)^{-1},$$

$$bias_{ATT,S}^2(\delta) = \mathbf{b}_S' \delta_{S^c} \delta_{S^c}' \mathbf{b}_S, \quad (3.3)$$
$$\mathbf{b}_S' = \text{ the first } |S^c| \text{ elements of the row vector in the bottom row of}$$
$$\left(\Lambda_S M^{ATT} \Lambda_S'\right)^{-1} \Lambda_S M^{ATT} \Lambda_{S^c}'.$$

By plugging in $\hat{M}^{ATT}$ and $\hat{\Sigma}^{ATT}$, we obtain consistent estimators for $\omega_{ATT,S}^2$ 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.

Let $\hat{\mathbf{t}}$ be a $|\mathcal{M}| \times 1$ column vector consisting of $\left\{\sqrt{n}\left(\hat{\tau}_S^{ATT} - \tau_n^{ATT}\right) : S \in \mathcal{M}\right\}$ and $\hat{\delta}_{\underline{S}^c} = \sqrt{n}\pi_{\underline{S}^c}(\hat{\gamma} - \gamma_0) = \sqrt{n}\hat{\gamma}_{\underline{S}^c}$, where $\pi_{\underline{S}^c}\gamma_0 = \mathbf{0}$ follows by Assumption DGP (iv). By noting that

the bias expression of (2.7) 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 $\left(\hat{\delta}_{\underline{S}^c}, \hat{\mathbf{t}}\right)$ as

$$\begin{pmatrix} \hat{\delta}_{\underline{S}^c} \\ \hat{\mathbf{t}} \end{pmatrix} \overset{P_{n,\delta}}{\rightsquigarrow} \begin{pmatrix} \Delta_{\underline{S}^c} \\ Z_\tau \end{pmatrix} \sim \mathcal{N}\left(\begin{pmatrix} \delta_{\underline{S}^c} \\ B\delta_{\underline{S}^c} \end{pmatrix}, \begin{pmatrix} \Omega_{11} & \Omega_{12} \\ \Omega_{21} & \Omega_{22} \end{pmatrix} \right), \tag{3.4}$$

where B is a $|\mathcal{M}| \times |\underline{S}^c|$ matrix, whose row vector corresponding to model S is $\mathbf{b}_S'\pi_{S^c}\pi_{\underline{S}^c}'$.[8] The covariance matrix $\Omega \equiv \begin{pmatrix} \Omega_{11} & \Omega_{12} \\ \Omega_{21} & \Omega_{22} \end{pmatrix}$ is the limit covariance matrix of

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

where T is a $|\mathcal{M}| \times (K+2)$ matrix with each row vector corresponding to model S being the bottom row vector of $-\left(\Lambda_S M^{ATT} \Lambda_S'\right)^{-1} \Lambda_S$. Accordingly, $\Omega_{11} = \pi_{\underline{S}^c}\mathcal{I}_\gamma^{-1}\pi_{\underline{S}^c}'$ is a submatrix of $\mathcal{I}_\gamma^{-1}$.

To establish optimality of averaging weights, define the following class of averaging weights that depend on data through $\hat{\delta}_{\underline{S}^c} = \sqrt{n}\hat{\gamma}_{\underline{S}^c}$ and $\left(\hat{B}, \hat{\Omega}\right)$, consistent estimators for (B, Ω):

$$\mathcal{C} \equiv \left\{ \hat{\mathbf{c}} = \mathbf{c}(\hat{\delta}_{\underline{S}^c}, \hat{B}, \hat{\Omega}) : \sum_{S \in \mathcal{M}} c_S(\hat{\delta}_{\underline{S}^c}, \hat{B}, \hat{\Omega}) = 1,\ \mathbf{c}(\cdot,\cdot,\cdot) \text{ is continuous a.e.} \right\}. \tag{3.5}$$

Note that $\mathcal{C}$ does not exhaust the universe of data-dependent averaging weights, since it excludes those that depend on data additionally through $\left(\hat{\tau}_S^{ATT} : S \in \mathcal{M}\right)$.[9] We suppress the second and third arguments of $\mathbf{c}(\hat{\delta}_{\underline{S}^c}, \hat{B}, \hat{\Omega})$ if the estimators $\left(\hat{B}, \hat{\Omega}\right)$ are replaced by the limiting true value (B, Ω), i.e., $\mathbf{c}(\hat{\delta}_{\underline{S}^c}) \equiv \mathbf{c}(\hat{\delta}_{\underline{S}^c}, B, \Omega)$. We consider the asymptotic trimmed mean-squared error as a performance criterion of averaging procedure $\hat{\mathbf{c}} \in \mathcal{C}$,

$$\begin{aligned} R_\infty(\hat{\mathbf{c}}, \delta_{\underline{S}^c}) &\equiv \lim_{\zeta\to\infty} \liminf_{n\to\infty} E_{P_{n,\delta}}\left[\min\left\{n(\hat{\tau}_{avg}^{ATT} - \tau_n^{ATT})^2, \zeta\right\}\right] \\ &= \lim_{\zeta\to\infty} \liminf_{n\to\infty} E_{P_{n,\delta}}\left[\min\left\{(\sqrt{n}\mathbf{c}(\hat{\delta}_{\underline{S}^c}, \hat{B}, \hat{\Omega})'\hat{\mathbf{t}})^2, \zeta\right\}\right], \end{aligned}$$

where the second argument $\delta_{\underline{S}^c}$ of $R_\infty(\cdot,\cdot)$ signifies that when $\hat{\mathbf{c}}$ is restricted to $\mathcal{C}$, the asymptotic MSE depends on the underlying data generating process only through the localization parameter $\delta_{\underline{S}^c}$.[10] The trimming is employed to circumvent the technical step of establishing uniform integrability of the sampling distribution of $n(\hat{\tau}_{avg}^{ATT} - \tau_n^{ATT})^2$. Next, we rank the performance of averaging

[8] The proof of Proposition 2.1 given in Appendix A yields the convergence in distribution of the joint distribution of $\hat{\delta}_{\underline{S}^c}$ and $\{\sqrt{n}(\hat{\tau}_S - \tau_n) : S \in \mathcal{M}\}$.

[9] Adopting the shrinkage estimators of the form considered in Hansen (2014b) to the current context, we can consider the weights that depend on data additionally through $(\hat{\tau}_S^{ATT} - \hat{\tau}^{ATT})$. Investigation of optimal averaging weights over a larger class of weights than $\mathcal{C}$ is out of scope of this paper.

[10] Our framework can be extended to different risk criteria such as the trimmed mean absolute deviation criterion. An advantage of the mean squared error criterion considered here is availability of a closed-form expression of the optimal averaging weights as shown below.

weights by a weighted average of the asymptotic MSEs with respect to a prior distribution for $\delta_{\underline{S}^c}$, $\mu(\delta_{\underline{S}^c})$,

$$R_{\infty}^{Bayes}(\hat{\mathbf{c}}) \equiv \int R_{\infty}(\hat{\mathbf{c}}, \delta_{\underline{S}^c}) d\mu(\delta_{\underline{S}^c}).$$

We hereafter refer to this criterion as Bayes asymptotic MSE.[11] Given true (B, Ω), let $\mathcal{C}(B, \Omega) \subset \mathcal{C}$ be the subset of averaging weights such that $\hat{\mathbf{c}} = \mathbf{c}(\hat{\delta}_{\underline{S}^c}, \hat{B}, \hat{\Omega})$ is continuous a.e. in $\hat{\delta}_{\underline{S}^c}$ when $(\hat{B}, \hat{\Omega})$ is set at true (B, Ω). Lemma A.3 in Appendix A shows that for $\hat{\mathbf{c}} \in \mathcal{C}(B, \Omega)$, the Bayes asymptotic risk can be expressed as

$$R_{\infty}^{Bayes}(\hat{\mathbf{c}}) \equiv \int E_{\Delta_{\underline{S}^c}|\delta_{\underline{S}^c}} \left[\mathbf{c}\left(\Delta_{\underline{S}^c}\right)' K\left(\Delta_{\underline{S}^c}, \delta_{\underline{S}^c}\right) \mathbf{c}\left(\Delta_{\underline{S}^c}\right) \right] d\mu(\delta_{\underline{S}^c}). \tag{3.6}$$

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

$$\begin{aligned} K\left(\Delta_{\underline{S}^c}, \delta_{\underline{S}^c}\right) &= \Omega_{22} - \Omega_{21}\Omega_{11}^{-1}\Omega_{12} \\ &+ \left(B - \Omega_{21}\Omega_{11}^{-1}\right)\left(\delta_{\underline{S}^c} - \Delta_{\underline{S}^c}\right)\left(\delta_{\underline{S}^c} - \Delta_{\underline{S}^c}\right)'\left(B - \Omega_{21}\Omega_{11}^{-1}\right)' \\ &+ \left(B - \Omega_{21}\Omega_{11}^{-1}\right)\left(\delta_{\underline{S}^c} - \Delta_{\underline{S}^c}\right)\Delta_{\underline{S}^c}'B' + B\Delta_{\underline{S}^c}\left(\delta_{\underline{S}^c} - \Delta_{\underline{S}^c}\right)'\left(B - \Omega_{21}\Omega_{11}^{-1}\right)' \\ &+ B\Delta_{\underline{S}^c}\Delta_{\underline{S}^c}'B'. \end{aligned} \tag{3.7}$$

Minimization of the Bayes asymptotic MSE (3.6) in $\mathbf{c}(\cdot)$ leads to the Bayes optimal averaging weights $\mathbf{c}^*(\cdot)$ in the limiting experiment, where the unknown object is $\delta_{\underline{S}^c}$ and $\Delta_{\underline{S}^c}$ serves as a sufficient statistic for it. Following the standard approach of limiting experiment analysis, we construct the finite sample analogue of $\mathbf{c}^*\left(\Delta_{\underline{S}^c}\right)$ by replacing the true (B, Ω) with their consistent estimators and $\Delta_{\underline{S}^c}$ with $\hat{\delta}_{\underline{S}^c}$. We hereafter refer to the procedure that uses the thus-constructed averaging weights as *Bayesian Limit Experiment (BayesLE) averaging*. The next proposition provides a closed-form expression of $\mathbf{c}^*\left(\Delta_{\underline{S}^c}\right)$ and shows that its finite sample analogue minimizes the Bayes asymptotic MSE.

Proposition 3.1 *Suppose Assumptions DGP and REG hold. Let $\mu\left(\delta_{\underline{S}^c}\right)$ be a proper prior, and let $K_{post}\left(\Delta_{\underline{S}^c}\right)$ be the posterior expectation of $K\left(\Delta_{\underline{S}^c}, \delta_{\underline{S}^c}\right)$ when $\Delta_{\underline{S}^c} \sim \mathcal{N}(\delta_{\underline{S}^c}, \Omega_{11})$,*

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

[11] Note that our definition of the Bayes risk in the limit experiment takes the average of the asymptotic risk instead of taking the limit of the average finite sample risk as considered in Hirano and Porter (2009) in the context of treatment choice.

(i) If $K_{post}\left(\Delta_{\underline{S}^c}\right)$ is nonsingular almost surely in $\Delta_{\underline{S}^c}$, the Bayes optimal model averaging weight in the limiting experiment, $\mathbf{c}^*(\cdot) \equiv \arg\min_{\mathbf{c}(\cdot)} \int E_{\Delta_{\underline{S}^c}|\delta_{\underline{S}^c}}\left[\mathbf{c}\left(\Delta_{\underline{S}^c}\right)' K\left(\Delta_{\underline{S}^c}, \delta_{\underline{S}^c}\right)\mathbf{c}\left(\Delta_{\underline{S}^c}\right)\right] d\mu(\delta_{\underline{S}^c})$, is unique almost surely in $\Delta_{\underline{S}^c}$, and is given by

$$\mathbf{c}^*\left(\Delta_{\underline{S}^c}\right) = \left[\mathbf{1}' K_{post}\left(\Delta_{\underline{S}^c}\right)^{-1}\mathbf{1}\right]^{-1}\left[K_{post}\left(\Delta_{\underline{S}^c}\right)^{-1}\mathbf{1}\right], \tag{3.8}$$

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

(ii) Let $\hat{K}\left(\Delta_{\underline{S}^c}, \delta_{\underline{S}^c}\right)$ be the sample analogue of $K\left(\Delta_{\underline{S}^c}, \delta_{\underline{S}^c}\right)$, where (B, Ω) is replaced by $\left(\hat{B}, \hat{\Omega}\right)$. Denote by $\hat{K}_{post}\left(\Delta_{\underline{S}^c}\right)$ the posterior expectation of $\hat{K}\left(\Delta_{\underline{S}^c}, \delta_{\underline{S}^c}\right)$ when the likelihood is $\Delta_{\underline{S}^c} \sim \mathcal{N}(\delta_{\underline{S}^c}, \hat{\Omega}_{11})$ and a prior for $\delta_{\underline{S}^c}$ is $\mu\left(\delta_{\underline{S}^c}\right)$. Then,

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

satisfies $R_\infty^{Bayes}(\hat{\mathbf{c}}) \geq R_\infty^{Bayes}\left(\mathbf{c}^*\left(\hat{\delta}_{\underline{S}^c}, \hat{B}, \hat{\Omega}\right)\right)$ for all $\hat{\mathbf{c}} \in \mathcal{C}(B, \Omega)$.

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, $\delta_{\underline{S}^c}|\Delta_{\underline{S}^c} \sim \mathcal{N}\left(\overline{\delta}_{\underline{S}^c}, \left(\Omega_{11}^{-1} + \Phi^{-1}\right)^{-1}\right)$, yields

$$\begin{aligned} K_{post}\left(\Delta_{\underline{S}^c}\right) &= \Omega_{22} - \Omega_{21}\Omega_{11}^{-1}\Omega_{12} \\ &\quad + \left[\left(B - \Omega_{21}\Omega_{11}^{-1}\right)\overline{\delta}_{\underline{S}^c} + \Omega_{21}\Omega_{11}^{-1}\Delta_{\underline{S}^c}\right]\left[\left(B - \Omega_{21}\Omega_{11}^{-1}\right)\overline{\delta}_{\underline{S}^c} + \Omega_{21}\Omega_{11}^{-1}\Delta_{\underline{S}^c}\right]' \\ &\quad + \left(B - \Omega_{21}\Omega_{11}^{-1}\right)\left(\Omega_{11}^{-1} + \Phi^{-1}\right)^{-1}\left(B - \Omega_{21}\Omega_{11}^{-1}\right)'. \end{aligned} \tag{3.9}$$

By plugging in $\hat{B}$ and $\hat{\Omega}$ and replacing $\Delta_{\underline{S}^c}$ by $\hat{\delta}_{\underline{S}^c}$, we obtain $\hat{K}_{post}\left(\hat{\delta}_{\underline{S}^c}\right)$ and the formula of $\mathbf{c}^*\left(\hat{\delta}_{\underline{S}^c}, \hat{B}, \hat{\Omega}\right)$ shown Proposition 3.1 (ii) computes the averaging weights that minimizes the Bayes asymptotic MSE.

The main reason that Proposition 3.1 assumes a proper prior is to guarantee that the Bayes asymptotic MSE is finite. In practice, requiring 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 if 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 (the Jeffreys prior for Gaussian means), then $K_{post}\left(\Delta_{\underline{S}^c}\right)$ is still well defined.

$$K_{post}\left(\Delta_{\underline{S}^c}\right) = \Omega_{22} - \Omega_{21}\Omega_{11}^{-1}\Omega_{12} + \left(B - \Omega_{21}\Omega_{11}^{-1}\right)\Omega_{11}\left(B - \Omega_{21}\Omega_{11}^{-1}\right)' + B\Delta_{\underline{S}^c}\Delta_{\underline{S}^c}'B', \tag{3.10}$$

Furthermore, the posterior risk has a well defined minimizer, given by (3.8), even if the resulting

Bayes asymptotic MSE (3.6) is unbounded.[12] We recommend using the uniform prior, unless the user has a strong prior opinion about the value of δ for the covariates. In our Monte Carlo studies and empirical application, we examine performance of the BayesLE-averaging estimator with the uniform prior.

Remark 3.1 *Hjort and Claeskens (2003, Sec. 5.4) propose the following way of obtaining weights. Given $\delta_{\underline{S}^c}$ and weight vector $\mathbf{c}$, the asymptotic MSE of the averaging estimator is written as $\mathbf{c}' E_{\Delta_{\underline{S}^c}|\delta_{\underline{S}^c}}\left[K\left(\Delta_{\underline{S}^c}, \delta_{\underline{S}^c}\right)\right]\mathbf{c} = \mathbf{c}'\left(\Omega_{22} - B\delta_{\underline{S}^c}\delta'_{\underline{S}^c}B'\right)\mathbf{c}$. The weights proposed by Hjort and Claeskens minimize the asymptotically unbiased estimator of the MSE in the limiting experiment,*

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

where $\Delta_{\underline{S}^c}\Delta'_{\underline{S}^c} - \Omega_{11}$ is an unbiased estimator for $\delta_{\underline{S}^c}\delta'_{\underline{S}^c}$. The solution to this minimization problem is given by

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

Note that $\mathbf{c}_{HC}\left(\Delta_{\underline{S}^c}\right)$ can be shown to differ from the BayesLE-averaging weights resulting from (3.9) for any of the conjugate normal priors as well as the weights corresponding to the uniform prior.[13]

Remark 3.2 *Model selection is a special case of averaging where the feasible weights are restricted to stepwise constant functions with their range restricted to $\left\{\mathbf{e}_1, \ldots, \mathbf{e}_{|\mathcal{M}|}\right\}$, where $\mathbf{e}_m$, $m = 1, \ldots, |\mathcal{M}|$, is the m-th column vector of $|\mathcal{M}| \times |\mathcal{M}|$ identity matrix. Let us denote a class of model selection procedures that select a set of covariates on the basis of $(\hat{\delta}_{\underline{S}^c}, \hat{B}, \hat{\Omega})$ by $\mathcal{C}_{sel} = \left\{\hat{\mathbf{c}} = \mathbf{c}(\hat{\delta}_{\underline{S}^c}, \hat{B}, \hat{\Omega}) : \mathbf{c}(\hat{\delta}_{\underline{S}^c}, \hat{B}, \hat{\Omega}) \in \left\{\mathbf{e}_1, \ldots, \mathbf{e}_{|\mathcal{M}|}\right\} \text{ for all } (\hat{\delta}_{\underline{S}^c}, \hat{B}, \hat{\Omega})\right\}$. By noting that with non-singular $\hat{K}_{post}\left(\hat{\delta}_{\underline{S}^c}\right)$, $\mathbf{c}^*\left(\hat{\delta}_{\underline{S}^c}, \hat{B}, \hat{\Omega}\right)$ derived in Proposition 3.1 is unique and never takes a corner solution, we can conclude that the optimal model averaging obtained in Proposition 3.1 strictly outperforms any of the model selection procedure in $\mathcal{C}_{sel} \cap \mathcal{C}(B, \Omega)$ in terms of Bayes asymptotic MSE.*

[12] One way to justify this averaging scheme would be to claim that the averaging weights corresponding to the uniform prior are obtained by a limit of the Bayes optimal weights with respect to a sequence of proper priors. Specifically, by noting that $K_{post}\left(\Delta_{\underline{S}^c}\right)$ of (3.9) converges to (3.10) as the prior variance matrix diverges to infinity, the optimal averaging weights under the uniform prior can be obtained as the limit of the Bayes optimal weights along a sequence of conjugate priors with diverging prior variances.

[13] Establishing the existence of a prior for $\delta_{\underline{S}^c}$ that supports $\mathbf{c}_{HC}(\Delta_{\underline{S}^c})$ as Bayes optimal in the limit experiment is left for future research.

Corollary 3.1 *Under the assumptions of Proposition 3.1,* $R_\infty^{Bayes}(\hat{\mathbf{c}}) > R_\infty^{Bayes}\left(\mathbf{c}^*\left(\hat{\delta}_{\underline{S}^c}, \hat{B}, \hat{\Omega}\right)\right)$ *holds for any* $\hat{\mathbf{c}} \in \mathcal{C}_{sel} \cap \mathcal{C}(B, \Omega)$.

The Monte Carlo studies shown in Section 4 below compares the MSE performances of BayesLE-averaging and the model selection procedure that selects a set of covariates based on the Bayes asymptotic risk. We find the MSE comparisons are consistent with the theoretical prediction of this corollary.

3.2 Post-averaging Inference

The optimality argument of BayesLE-averaging proposed in Proposition 3.1 concerns point estimation and has little to say about how to proceed to interval estimation. This section presents a construction of confidence intervals based on the sampling distribution of the averaging estimator by adopting the two-stage confidence procedure proposed by Claeskens and Hjort (2008). The proposed confidence intervals guarantee nominal coverage, although their coverage probability can be conservative.

Let $(1-\beta) \in (0,1)$ be a nominal coverage probability and let $\beta_1, \beta_2 > 0$ satisfy $\beta_1 + \beta_2 = \beta$. Given a value of localization parameter $\delta_{\underline{S}^c}$, the weak convergence of $\sqrt{n}(\hat{\delta}'_{\underline{S}^c}, \hat{\mathbf{t}}')'$ shown in (3.4) implies that the averaging estimator of Proposition 3.1 converges to $\sqrt{n}(\hat{\tau}_{avg}^{ATT} - \tau_n^{ATT}) \overset{P_{n,\delta}}{\rightsquigarrow} \mathbf{c}^*(\Delta_{\underline{S}^c}) Z_\tau$. Based on this asymptotic distribution, let $CI_{1-\beta_1}^{ATT}(\Delta_{\underline{S}^c}, Z_\tau | \delta_{\underline{S}^c})$ be an interval estimator for ATT that satisfies $\Pr\left(\tau_0^{ATT} \in CI_{1-\beta_1}^{ATT}(\Delta_{\underline{S}^c}, Z_\tau | \delta_{\underline{S}^c})\right) = 1 - \beta_1$. Since random variable $\mathbf{c}^*(\Delta_{\underline{S}^c}) Z_\tau$ is easy to simulate, it is straightforward to numerically approximate $CI_{1-\beta_1}^{ATT}(\Delta_{\underline{S}^c}, Z_\tau | \delta_{\underline{S}^c})$.

The two step confidence procedure proceeds as follows. In the first step, we construct a confidence set (ellipsoid) for $\delta_{\underline{S}^c}$ with confidence level $(1-\beta_2)$ by inverting the likelihood ratio test,

$$CS_{1-\beta_2} \equiv \left\{ \delta_{\underline{S}^c} : (\hat{\delta}_{\underline{S}^c} - \delta_{\underline{S}^c})' \hat{\Omega}_{11}^{-1} (\hat{\delta}_{\underline{S}^c} - \delta_{\underline{S}^c}) \leq \chi^2_{1-\beta_2}(dim(\delta_{\underline{S}^c})) \right\},$$

where $\chi^2_{1-\beta_2}(dim(\delta_{\underline{S}^c}))$ is the $(1-\beta_2)$-th quantile of the χ^2-statistic with degree of freedom equal to the dimension of $\delta_{\underline{S}^c}$. In the second step, we construct a confidence interval for ATT, $CI_{1-\beta}^{ATT}(\hat{\delta}_{\underline{S}^c}, \hat{\mathbf{t}})$, by taking the union of $CI_{1-\beta_1}^{ATT}(\hat{\delta}_{\underline{S}^c}, \hat{\mathbf{t}} | \delta_{\underline{S}^c})$ over $\delta_{\underline{S}^c} \in CS_{1-\beta_2}$. It can be shown that the asymptotic coverage probability of $CI_{1-\beta}^{ATT}(\hat{\delta}_{\underline{S}^c}, \hat{\mathbf{t}})$ is bounded from below by $1-\beta$ irrespective of the value of δ, and hence the confidence intervals for ATT are asymptotically uniformly valid at least over the class of propensity scores that meet Assumptions DGP (i)-(ii) and REG. See Appendix A for a proof of these claims. In the empirical application presented below, we implement this two-step procedure by taking the union of $CI_{1-\beta_1}^{ATT}(\hat{\delta}_{\underline{S}^c}, \hat{\mathbf{t}} | \delta_{\underline{S}^c})$ over randomly sampled values of $\delta_{\underline{S}^c} \in CS_{1-\beta_2}$.

Parameter	Description	Benchmark value
n	Number of observations	100
K	Number of regressors	3
β_1	Outcome equation coefficient of X_1	0.5
β_2	Outcome equation coefficient of X_k, $k > 1$	$0.5/(K-1)$
γ	Selection equation coefficient	1
σ_u	Conditional st. dev. outcome equation	2

Table 1: Parameters for the simulations in Section 4, and their benchmark values.

4 Monte Carlo Study

In this section, we perform a simulation experiment to study the behavior of the averaging estimator proposed in Section 3. We show that a bias-variance trade-off exists between a small and a large models for the NPW-ATT estimator, and find MSE gains for the model averaging estimator.

We will use a model with treatment outcome $Y(1) = u_1$, control outcome

$$Y(0) = -\beta_1 X_1 - \beta_2 \sum_{k=2}^{K} X_k + u_0,$$

and selection equation $P(D = 1|\, X) = G\left(\frac{\gamma}{K}\sum_{k=1}^{K} X_k\right)$, where G is the logistic function. The outcome equation error terms (u_0, u_1) are generated from a zero mean normal distribution. The regressors are generated, independently of those error terms, from a multivariate normal distribution with mean 0, standard deviation 1, and pairwise covariance 0.5.

The design parameters and their benchmark values are listed in Table 4. We let the first regressor, X_1, be more important than the remaining regressors by letting its regression coefficient β_1 be larger than the coefficient of each of the remaining regressors, β_2. We have normalized the sum of the regression coefficients to 1, so that the covariate X_1 accounts for a share β_1 of the model, and the other regressors share the remaining $1 - \beta_1$ equally. In the benchmark design, each of the regressors (X_2, X_3) are only half as important as the first one. As a result, the first regressor X_1 is very important, and should probably be included in estimation, but there may be some advantage from leaving out X_2 or X_3.

Note that the parameters n, K, and γ affect the bias-variance trade-off. Increasing the value of K increases the number of coefficients that have to be estimated, but reduces the bias of leaving out a single regressor since the coefficient for each regressor, $\beta_2 = (1 - \beta_1)\,/\,(K - 1)$, decreases in K. The selection equation coefficient γ controls the strength of the selection effect, which is assumed

to be the same for all regressors. Increasing γ increases the bias of leaving out a regressor, and affects regressor overlap. We investigate the role of these parameters in detail in the sensitivity analysis below.

For our simulation design, the average treatment effect is 0. By using the properties of our design, it can be shown that the average treatment effect on the treated $E\left(Y\left(1\right)-Y\left(0\right)|\,D=1\right)=E\left(X|\,D=1\right)\beta$ does not depend on the design parameters (n,β_1,σ_u) but depends on the number of regressors, K, and on γ, which governs the relationship between the regressors and the treatment indicator.

The model averaging estimator depends on estimators of the matrices B and Ω in equation (3.4). Estimators for B and Ω are obtained from the full model using sample analog estimators that were shown to be consistent in Appendix A. Note that the different submodel estimators are highly correlated. Therefore, the inversion of K_{post} can be problematic. For this reason, we will regularize K_{post} before inversion, using the approach in Carrasco et al. (2007). Results for each model are based on 10000 replications.

We will refer to the model with all regressors as the "full model", and tothe model that only includes X_1 and a constant term as the "small model". On top of the submodel estimators, we report the following three estimators: (1) the infeasible "Best submodel" estimator, which is the submodel estimator with the lowest MSE across simulations; (2) the "BayesLE-averaging" estimator with improper uniform $\mu(\delta_{\underline{S}^c})$ based on all 2^K-1 or $2^{K-1}-1$ submodel estimators; and (3) the "Selection" estimator, which chooses the estimator with the lowest estimated MSE.[14]

Results for the benchmark simulation design. The results for the benchmark simulations can be found in Table 4. Given a number of regressors K, we either consider the $2^{K-1}-1$ submodels that include a constant term and the important regressor X_1, or we consider all 2^K-1 submodels. The former corresponds to the more realistic situation that a researcher has some idea about what the important regressors are, but is unsure about including a number of less important control regressors.

Several findings are worth noting. First, note that all the estimators that leave out the relevant regressor X_1 are severely biased due to omitting the important regressor. Second, there is a clear bias-variance trade-off: the small model (only X_1) outperforms the full model (all regressors). Third, the full model estimator has the lowest bias. Fourth, the BayesLE-averaging estimator seems to have the best overall performance in terms of MSE. In particular, it outperforms the selection estimator, and it achieves the MSE of the best submodel. Finally, the performance of the selection procedure deteriorates slightly by the inclusion of poorly performing models (i.e. models

[14]The selection estimator is obtained by solving $\min_{\mathbf{c}\in\mathcal{C}_{sel}}\mathbf{c}'E_{\delta_{\underline{S}^c}|\Delta_{\underline{S}^c}}\left[\hat{K}\left(\Delta_{\underline{S}^c},\delta_{\underline{S}^c}\right)|\Delta_{\underline{S}^c}=\hat{\delta}_{\underline{S}^c}\right]\mathbf{c}$, where $\hat{K}\left(\Delta_{\underline{S}^c},\delta_{\underline{S}^c}\right)$ is as defined in Proposition 3.1 (ii).

	Submodels with X_1			All submodels		
Estimator	Bias	SD	MSE	Bias	SD	MSE
Small:$\{X_1\}$	-0.162	0.471	0.249	-0.162	0.471	0.249
$\{X_1, X_2\}$	-0.065	0.502	0.256	-0.065	0.502	0.256
Full:$\{X_1, X_2, X_3\}$	-0.016	0.522	0.273	-0.016	0.522	0.273
$\{X_1, X_3\}$	-0.064	0.501	0.255	-0.064	0.501	0.255
$\{X_2\}$	N/A	N/A	N/A	-0.244	0.472	0.282
$\{X_2, X_3\}$	N/A	N/A	N/A	-0.121	0.502	0.267
$\{X_3\}$	N/A	N/A	N/A	-0.244	0.474	0.284
Best submodel	-0.162	0.471	0.249	-0.162	0.471	0.249
BayesLE-averaging	-0.075	0.493	0.249	-0.125	0.481	0.247
Selection	-0.036	0.524	0.275	-0.055	0.526	0.280

Table 2: Simulation results for the benchmark setup.

without X_1), whereas including these poorly performing models leads to a slight improvement the performance of the averaging estimator. The results in Table 4 suggest that the averaging procedure is robust against the inclusion of poorly performing models.

Sensitivity analysis. We now conduct a sensitivity analysis to check whether the conclusions from the simulation results are robust to changes in the design parameters, and to investigate the role of regressor overlap. The results are presented in Figures 1 and 2. Unless otherwise mentioned, we fix parameter values to their benchmark values in Table 4. We let $n = 100$ (left column) and $n = 300$ (right column), and we let $K = 3$ (top row) and $K = 6$ (bottom row). For each scenario, we plot the results as a function of γ, the regression coefficient in the selection equation.[15]

We report results for the full model estimator (based on all covariates), for the small model estimator (based on X_1 only) and for two BayesLE-averaging estimators. The first one ("All") is based on all $2^K - 1$ submodel estimators that include X_1. The second one ("Nested") combines estimators from nested models only. By a nested model, we refer to a model with combinations of regressors that can include X_k only if they include X_{k-1}. For example, for the case $K = 3$, the researcher considers three submodel estimators: one based on including X_1; one based on including X_1 and X_2; and one that uses all regressors. We use 20000 draws for each set of simulation design parameter values.

We first consider the bias for the estimators (Figure 1). The solid line corresponds to the true

[15] We evaluate results at $\gamma \in \{0, 0.1, 0.2, \cdots, 1.9, 2\}$. For values of $\gamma > 2$, overlap becomes so poor (see Table 3) that alternative estimation procedures should be considered.

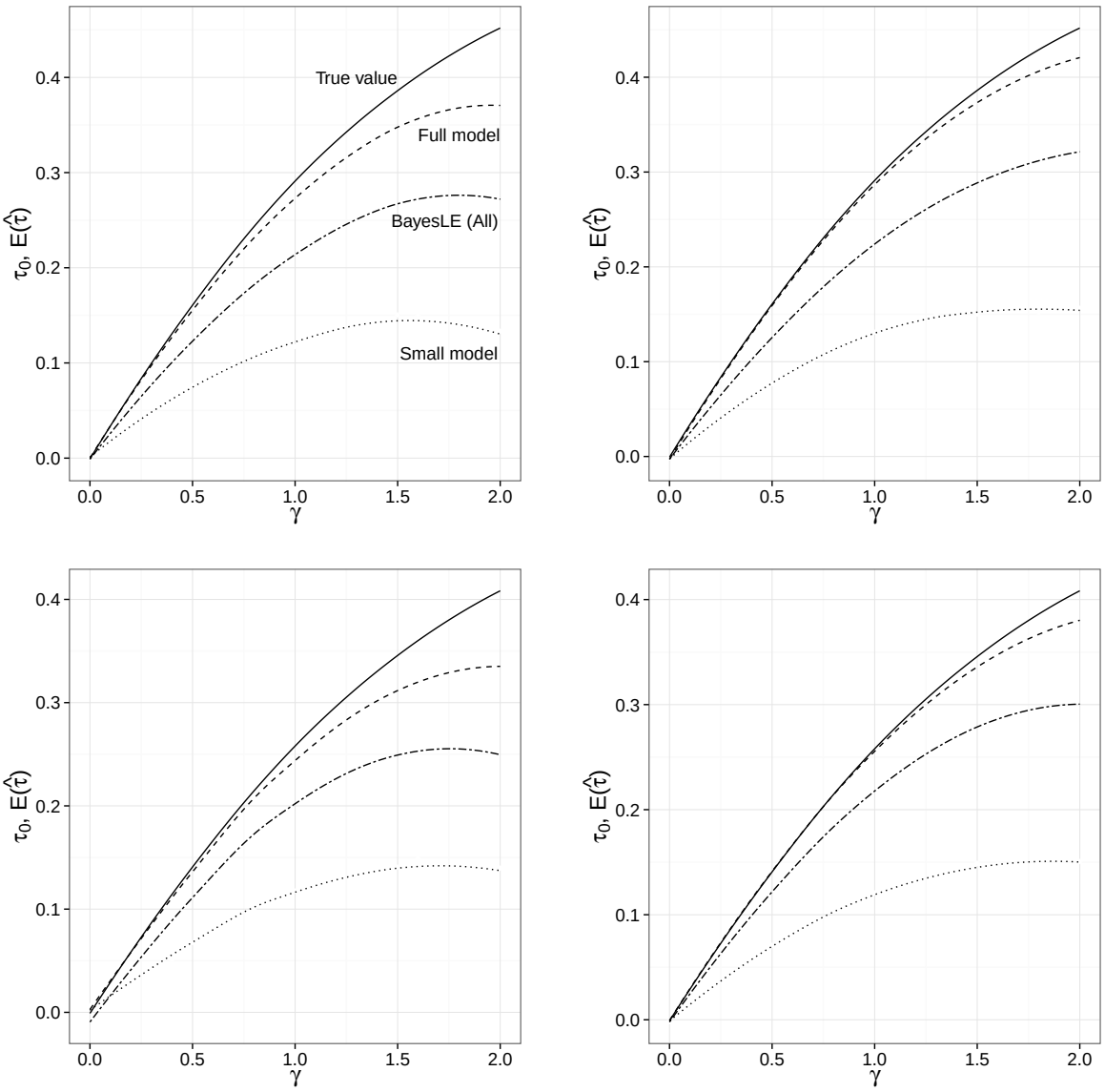

Figure 1: Results for sensitivity analysis: bias. Solid line corresponds to the true value of the ATT, τ_0. We plot the simulated expected value of three estimators. Left column $n = 100$; right column: $n = 300$. Top row: $K = 3$; bottom row: $K = 6$.

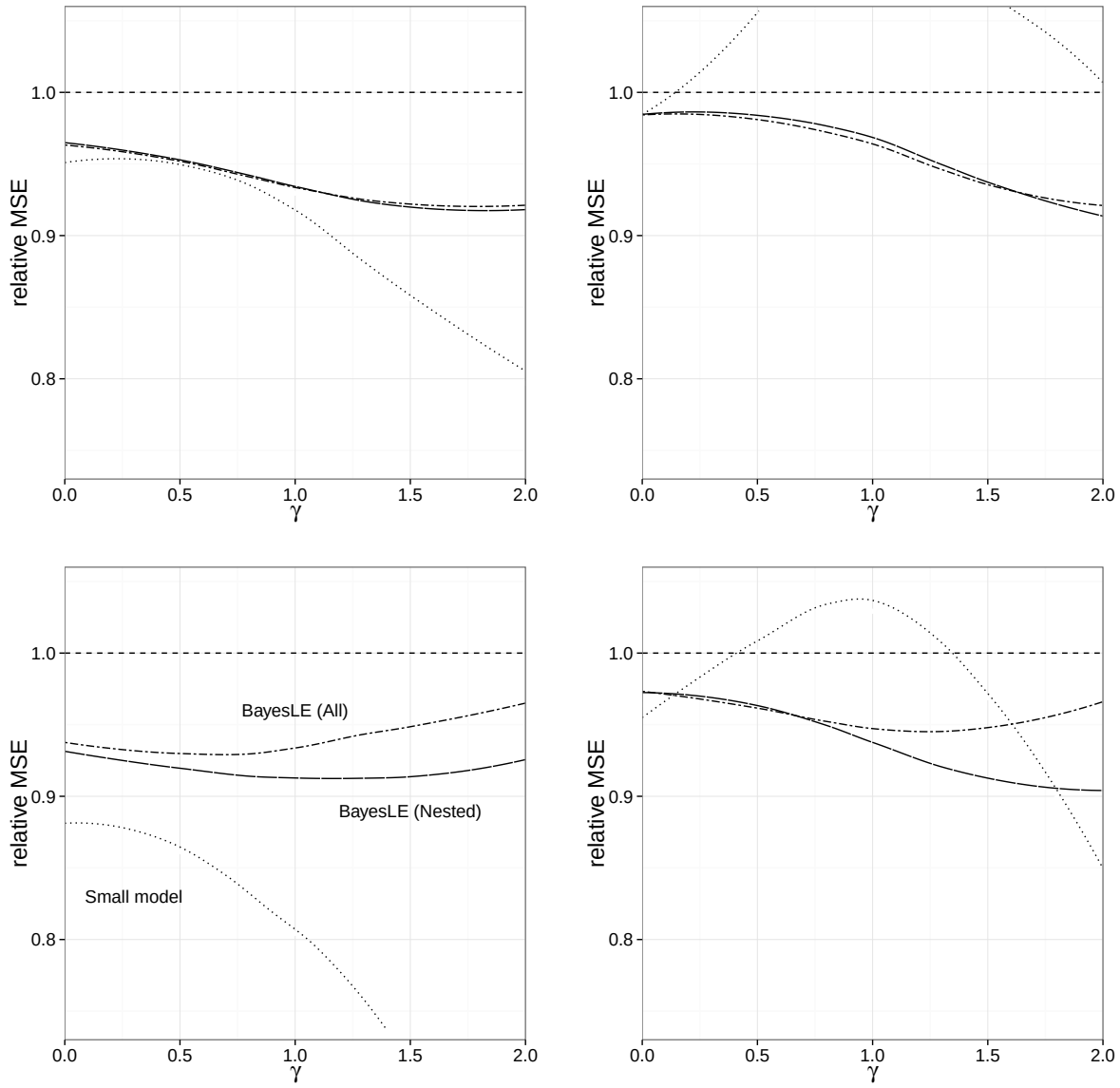

Figure 2: Results for sensitivity analysis: mean squared error. The mean squared error is relative to that of the full model (dashed). We plot the simulated relative mean squared error for three estimators. Left column $n = 100$; right column: $n = 300$. Top row: $K = 3$; bottom row: $K = 6$.

K	γ	$P\left(p\left(X\right)\geq 0.9\middle\vert D=1\right)$	$P\left(p\left(X\right)\geq 0.95\middle\vert D=1\right)$	$P\left(p\left(X\right)\geq 0.99\middle\vert D=1\right)$
3	0.5	0.131	0.038	0.001
3	1	0.476	0.332	0.113
3	2	0.732	0.645	0.456
6	0.5	0.388	0.250	0.064
6	1	0.669	0.573	0.375
6	2	0.832	0.800	0.664

Table 3: Probability of exceeding certain propensity score values for the treatment group, for various values of K and γ.

value of the ATT, which is increasing in γ . First, note that the full model estimator (dashed line) is not unbiased. Comparing the left column ($n = 100$) to the right column ($n = 300$), suggests that this is a finite sample bias. The bias is increasing in γ, which is likely to be a result of the decrease in regressor overlap (see Table 3). Second, note that the bias for the small model is always bigger than the bias of the full model estimator. The bias of the BayesLE-averaging ("All") procedure is in between that of the small and full model estimators. We do not present the bias for BayesLE-averaging ("Nested"), as it is very similar to that of BayesLE-averaging ("All").

Next, we consider the relative mean squared error of the small model estimator and the BayesLE-averaging estimators relative to the full model estimator (Figure 2). First, note that the BayesLE-averaging procedures outperforms the full model estimator for the full range of the parameter space considered in our simulations. Second, note that the relative MSE of the small model estimator is non-monotonic. This is related to the two effects that changing γ has in our simulation design. Increasing γ (i) increases the bias of leaving out regressors, (ii) decreases overlap, which makes it more favorable to consider subsets of regressors. Third, note that increasing the number of regressors improves the relative performance of the small model estimator and for the BayesLE-averaging estimators. Increasing the number of observations decreases the relative performance. This is not surprising, because increasing n effectively changes the value of the misspecification parameter δ. Finally, we point out that the value of the mean squared error of the full model estimator is monotonically increasing in γ (not shown in Figure 2).

5 Empirical application

In this section, we apply the methods discussed in Sections 2 and 3 to the data set analyzed in LaLonde (1986) and Dehejia and Wahba (1999). In the context of model selection with l_1-

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 ($)	No	2096	14017
re75	1975 earnings ($)	Yes	1532	13651
hispanic	1 if hispanic	No	0.06	0.07
married	1 if married	Yes	0.19	0.71
age^2	-	Yes		
$re75^2$	-	No		
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.

penalty, this data set is also analyzed by Farrell (2013). 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 benefited from work experience, and some counseling. 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 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.[16] 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.[17] 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

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

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

	Full model		Small model	
Variable		SE	$\hat{\gamma}$	SE
age	0.64	0.64	0.62	0.08
education	−0.19	0.03	−0.21	0.03
black	3.99	0.26	3.65	0.21
hispanic	1.59	0.41		
married	−1.40	0.24	−1.39	0.23
re74	−0.07	0.03		
re75	−0.29	0.06	−0.27	0.03
$\text{age}^2/100$	−1.06	0.14	−1.04	0.14
re75^2	1.52	0.87		
n		14559		14559

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 estimate is multiplied by the standard deviation of the regressor.

adjustment for age, education, and race.[18] 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 order terms and to excluding 1974 earnings.

We consider the set of variables and transformations in Table 4. The treatment and control groups have sizable differences in terms of their observable characteristics, so a difference in means is unlikely to be unbiased for the average treatment effect. We consider a scenario with 8 submodels: for each variable in (hispanic, married, re75^2), we are unsure whether it should be included in the propensity score. The other six variables are always included. We use a logit form for the selection equation. Finally, we trim the 10% of observations with the lowest estimated propensity scores.

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 precisely estimated in the small model.

Table 6 reports 90% confidence intervals for the experimental estimate, the full model estimate, and the BayesLE-averaging estimate. For the BayesLE estimator, we use the two-step confidence

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

Method	Estimate	SE	90%-CI
Experimental	1672	637	[627, 2717]
Full model	1358	753	[123, 2593]
Bayes	1468	-	[110, 2873]

Table 6: Estimates and confidence intervals for three procedures.

procedure described in Section 3.2 with $\beta_1 = \beta_2 = 0.05$. All confidence intervals are quite wide, which is consistent with the findings in LaLonde and DW. Post-averaging inference leads to less precise inference than using standard inference 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 MSE performance. The procedure we use is known to be conservative (Hjort and Claeskens, 2008, p. 211). A promising development for improving this is Liu (2013).

6 Concluding Remarks

We proposed a model averaging procedure for normalized propensity score weighted estimation of the ATT by extending the framework of the focused information criterion and frequentist model averaging to the semiparametric estimation of ATT. The aim of these procedures is to construct the most accurate estimator for ATT in terms of MSE, under the assumption that unconfoundedness holds and that the propensity scores are correctly specified in a most complicated specification provided by the user. The resulting procedure is easy to implement, and can offer a reference estimate of the ATT in the presence of the uncertainty in propensity score specifications. Our Monte Carlo evidence shows that the proposed procedure enjoys good MSE improvement compared to post-model selection estimator as well as the estimators constructed in the 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 remain out of the scope of this paper. 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. Third, this paper mainly focusses on point estimation, and relies on existing idea to construct conservative confidence intervals. It would be interesting to develop theory for the construction of less conservative post-averaging inference. We leave these important issues for future research.

Appendix

A Lemmas and Proofs

Following Busso et al. (2014), we formulate the NPW estimations for ATE and ATT by the following system of just-identified moment conditions:

$$E_{P_{n,\delta}}\left[\mathbf{m}^{ATE}\left(Z_i,\theta_n^{ATE}\right)\right]=E_{P_{n,\delta}}\begin{pmatrix}\frac{\left(D_i-G\left(W_i'\gamma_n\right)\right)}{G\left(W_i'\gamma_n\right)\left[1-G\left(W_i'\gamma_n\right)\right]}g\left(W_i'\gamma_n\right)W_i\\ \left[\frac{D_i}{G(W_i'\gamma_n)}+\frac{1-D_i}{1-G\left(W_i'\gamma_n\right)}\right]\left(Y_i-\tau_0^{ATE}D_i-\mu_0\right)\\ \left[\frac{D_i}{G(W_i'\gamma_n)}+\frac{1-D_i}{1-G\left(W_i'\gamma_n\right)}\right]\left(Y_i-\tau_0^{ATE}D_i-\mu_0\right)D_i\end{pmatrix}=\mathbf{0}$$

$$E_{P_{n,\delta}}\left[\mathbf{m}^{ATT}\left(Z_i,\theta_n^{ATT}\right)\right]=E_{P_{n,\delta}}\begin{pmatrix}\frac{\left(D_i-G\left(W_i'\gamma_n\right)\right)}{G\left(W_i'\gamma_n\right)\left[1-G\left(W_i'\gamma_n\right)\right]}g\left(W_i'\gamma_n\right)W_i\\ \left[D_i+(1-D_i)\left(\frac{G\left(W_i'\gamma_n\right)}{1-G\left(W_i'\gamma_n\right)}\right)\right]\left(Y_i-\tau_n^{ATT}D_i-\alpha_n\right)\\ \left[D_i+(1-D_i)\left(\frac{G\left(W_i'\gamma_n\right)}{1-G\left(W_i'\gamma_n\right)}\right)\right]\left(Y_i-\tau_n^{ATT}D_i-\alpha_n\right)D_i\end{pmatrix}=\mathbf{0}. \tag{A.1}$$

where $Z_i\equiv(Y_i,D_i,W(X_i))$ is a random vector of an observation whose probability law is induced by $P_{n,\delta}$ defined in (2.1), and $\theta_n^{ATE}\equiv\left(\gamma_n,\mu_0,\tau_0^{ATE}\right)'\in\mathbb{R}^{K+2}$ and $\theta_n^{ATT}\equiv\left(\gamma_n,\alpha_n,\tau_n^{ATT}\right)'\in\mathbb{R}^{K+2}$ are the parameter vectors solving the population moment conditions for the ATE and ATT, respectively. Note that parameters μ_0 and τ_0^{ATE} in θ_n^{ATE} do not depend on n, since the distribution of potential outcomes do not drift with n. The first K elements of the moment vectors $\mathbf{m}^{ATE}\left(Z_i,\theta_n^{ATE}\right)$ and $\mathbf{m}^{ATT}\left(Z_i,\theta_n^{ATT}\right)$ are the score vector from the propensity score estimation and are common between the ATE and ATT moment conditions. The sample analogue of these moment conditions yields the NPW estimators (2.2) and (2.3) in the largest model.

Let $\theta_0^{ATE}\equiv\left(\gamma_0',\mu_0,\tau_0^{ATE}\right)'$ and $\theta_0^{ATT}\equiv\left(\gamma_0',\alpha_0,\tau_0^{ATT}\right)'$. We denote by $\hat{\theta}^{ATE}=\left(\hat{\gamma}',\hat{\mu},\hat{\tau}^{ATE}\right)'$ and $\hat{\theta}^{ATT}=\left(\hat{\gamma}',\hat{\alpha},\hat{\tau}^{ATT}\right)'$ the method of moment estimators in the largest model. For each selection of covariates $S\in\mathcal{M}$, we define

$$\begin{aligned}\gamma^S&=\pi_S'\pi_S\gamma+\left(I-\pi_S'\pi_S\right)\gamma_0\\ \theta^{ATE,S}&=\left(\gamma^{S\prime},\mu,\tau^{ATE}\right)',\ \theta^{ATT,S}=\left(\gamma^{S\prime},\alpha,\tau^{ATT}\right)'\end{aligned}$$

where π_S is the selection matrix defined in the main text. γ^S is a $(K\times1)$ vector obtained by replacing the elements of γ that are not included in S with their benchmark values γ_0 (zeros by Assumption DGP (iv)). In particular, for a sequence of DGPs $\{P_{n,\delta}\}$ satisfying Assumption DGP,

we define

$$\gamma_n^S = \pi_S'\pi_S\gamma_n + \left(I - \pi_S'\pi_S\right)\gamma_0,$$
$$\theta_n^{ATE,S} = \left(\gamma_n^{S\prime}, \mu_0, \tau_0^{ATE}\right)', \quad \theta_n^{ATT,S} = \left(\gamma_n^{S\prime}, \alpha_n, \tau_n^{ATT}\right)'.$$

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, define a $(K \times 1)$ vector

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

Let $\mathbf{m}_n^{ATE}(\theta) = \frac{1}{n}\sum_{i=1}^n \mathbf{m}^{ATE}(Z_i, \theta^{ATE})$ and $\mathbf{m}_n^{ATT}(\theta) = \frac{1}{n}\sum_{i=1}^n \mathbf{m}^{ATT}(Z_i, \theta^{ATT})$. Using Λ_S defined in the main text, the NPW estimators in model S solve the following $(|S|+2)$-dimensional just-identifying sample moments,

$$\Lambda_S \mathbf{m}_n^{ATE}(\hat{\theta}^{ATE,S}) = \mathbf{0},$$

$$\Lambda_S \mathbf{m}_n^{ATT}(\hat{\theta}^{ATT,S}) = \mathbf{0},$$

with

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

where $\hat{\tau}_S^{ATE}$ and $\hat{\tau}_S^{ATT}$ are the NPW estimators for ATE and ATT in model S shown in the main text, and $\hat{\mu}_S$ and $\hat{\alpha}_S$ are the corresponding value of μ and α solving the moment conditions in model S.

We first show a basic lemma that extends Lemma 4.3 of Newey and McFadden to a triangular array of random variables involving estimated parameters.

Lemma A.1 *Let Z_i, $i = 1, \ldots, n$, be i.i.d sequence of random vectors following P_n, $n = 1, 2, \ldots$. Let $\theta_n \equiv \theta(P_n)$ be a sequence of parameter vectors corresponding to P_n. Let $a(Z, \theta)$ be a real-valued function of an observation Z and parameter θ. Suppose $\hat{\theta}$ an estimator for θ satisfies $\left\|\hat{\theta} - \theta_n\right\| = o_{P_n}(1)$, and let $\{\epsilon_n\}$be a converging sequence that satisfies $P_n(\left\|\hat{\theta} - \theta_n\right\| \leq \epsilon_n) \to 1$ as $n \to \infty$. If (i) $E_{P_n}\left[\sup_{\|\theta - \theta_n\| \leq \epsilon_n} |a(Z, \theta) - a(Z, \theta_n)|\right] \to 0$ as $n \to \infty$, and (ii) there exists $\lambda > 0$ such that $E_{P_n}\left[|a(Z, \theta_n)|^{1+\lambda}\right] < \infty$, then*

$$\left|\frac{1}{n}\sum_{i=1}^n a(Z_i, \hat{\theta}) - E_{P_n}\left[a(Z, \theta_n)\right]\right| = o_{P_n}(1).$$

Proof. By the triangular inequality,

$$\left|\frac{1}{n}\sum_{i=1}^{n} a(Z_i,\hat{\theta}) - E_{P_n}\left[a(Z,\theta_n)\right]\right| \leq \left|\frac{1}{n}\sum_{i=1}^{n} a(Z_i,\hat{\theta}) - \frac{1}{n}\sum_{i=1}^{n} a(Z_i,\theta_n)\right| + \left|\frac{1}{n}\sum_{i=1}^{n} a(Z_i,\theta_n) - E_{P_n}\left[a(Z,\theta_n)\right]\right|. \tag{A.2}$$

To show the first term in the right hand side converges, let us define event $\Omega_n \equiv \left\{\left\|\hat{\theta} - \theta_n\right\| \leq \epsilon_n\right\}$ and random variable $\Delta_n(Z) = \sup_{\|\theta-\theta_n\|\leq\epsilon_n} |a(Z,\theta) - a(Z,\theta_n)|$. We then have for any $\eta > 0$,

$$\begin{aligned}
&P_n\left(\left|\frac{1}{n}\sum_{i=1}^{n} a(Z_i,\hat{\theta}) - \frac{1}{n}\sum_{i=1}^{n} a(Z_i,\theta_n)\right| > \eta\right) \\
\leq &P_n\left(\left\{\left|\frac{1}{n}\sum_{i=1}^{n} a(Z_i,\hat{\theta}) - \frac{1}{n}\sum_{i=1}^{n} a(Z_i,\theta_n)\right| > \eta\right\} \cap \Omega_n\right) + P_n(\Omega_n^c) \\
\leq &P_n\left(\left\{\frac{1}{n}\sum_{i=1}^{n} \Delta_n(Z_i) > \eta\right\}\right) + o(1) \\
\leq &E_{P_n}\left[\Delta_n(Z_i)\right]/\eta + o(1) \\
= &o(1),
\end{aligned}$$

where the second like uses $\left|\frac{1}{n}\sum_{i=1}^{n} a(Z_i,\hat{\theta}) - \frac{1}{n}\sum_{i=1}^{n} a(Z_i,\theta_n)\right| \leq \frac{1}{n}\sum_{i=1}^{n} \Delta_n(Z_i)$ on event Ω_n, the third line follows by the Markov inequality, and the last line follows from assumption (i).

Note that assumption (ii) implies $E_{P_n}\left[\left|a_k(Z,\theta_n) - E_{P_n}\left[a_k(Z,\theta_n)\right]\right|^{1+\lambda}\right] < \infty$. Hence, the law of large numbers for a triangular array of random variables (see e.g., Lemma 11.4.2 of Lehmann and Romano (2005)) yields $\left|\frac{1}{n}\sum_{i=1}^{n} a_k(Z_i,\theta_n) - E_{P_n}\left[a_k(Z,\theta_n)\right]\right| = o_{P_n}(1)$. Hence, the conclusion follows. ■

The next lemma collects consistency and asymptotic normality results in our local asymptotic analysis, which are useful to prove Proposition 2.1 and the claims given in Section 3 of the main text.

Lemma A.2 *Let $\{P_{n,\delta}\} \in \mathcal{P}$ be a sequence of data generating processes indexed by localization parameter δ. Under Assumptions DGP and REG in the main text, the following claims hold:*

(i) $\left\|\hat{\theta}^{ATE} - \theta_n^{ATE}\right\| = o_{P_{n,\delta}}(1)$ *and* $\left\|\hat{\theta}^{ATT} - \theta_n^{ATT}\right\| = o_{P_{n,\delta}}(1)$.

(ii) $\left\|\hat{\theta}^{ATE,S} - \theta_n^{ATE}\right\| = o_{P_{n,\delta}}(1)$ *and* $\left\|\hat{\theta}^{ATT,S} - \theta_n^{ATT}\right\| = o_{P_{n,\delta}}(1)$ *for every* $S \in \mathcal{M}$.

(iii) *Let* $M^{ATE} \equiv E_{P_0}\left[\frac{\partial}{\partial\theta'}\mathbf{m}^{ATE}\left(Z,\theta_0^{ATE}\right)\right]$ *and* $M^{ATT} \equiv E_{P_0}\left[\frac{\partial}{\partial\theta'}\mathbf{m}^{ATT}\left(Z,\theta_0^{ATT}\right)\right]$. *Let* $\bar{\theta}^{ATE}$ *and* $\bar{\theta}^{ATT}$ *be estimators for* θ^{ATE} *and* θ^{ATT} *that satisfy* $\left\|\bar{\theta}^{ATE} - \theta_n^{ATE}\right\| = o_{P_{n,\delta}}(1)$ *and*

$\left\|\bar{\theta}^{ATT} - \theta_n^{ATT}\right\| = o_{P_{n,\delta}}(1)$, *respectively. Then,*

$$\left\| \frac{1}{n}\sum_{i=1}^{n} \frac{\partial}{\partial \theta'} \mathbf{m}^{ATE}\left(Z_i, \bar{\theta}^{ATE}\right) - M^{ATE} \right\| = o_{P_{n,\delta}}(1),$$

$$\left\| \frac{1}{n}\sum_{i=1}^{n} \frac{\partial}{\partial \theta'} \mathbf{m}^{ATT}\left(Z_i, \bar{\theta}^{ATT}\right) - M^{ATT} \right\| = o_{P_{n,\delta}}(1),$$

(iv) *Denote the variance-covariance matrices of* $\mathbf{m}^{ATE}\left(Z_i, \theta_0^{ATE}\right)$ *and* $\mathbf{m}^{ATE}\left(Z_i, \theta_0^{ATE}\right)$ *by* Σ^{ATE} *and* Σ^{ATT}*, respectively. Let* $\bar{\theta}^{ATE}$ *and* $\bar{\theta}^{ATT}$ *be estimators for* θ^{ATE} *and* θ^{ATT} *as defined in (iii).*

$$\left\| \frac{1}{n}\sum_{i=1}^{n} \mathbf{m}^{ATE}\left(Z_i, \bar{\theta}^{ATE}\right) \mathbf{m}^{ATE}\left(Z_i, \bar{\theta}^{ATE}\right)' - \Sigma^{ATE} \right\| = o_{P_{n,\delta}}(1),$$

$$\left\| \frac{1}{n}\sum_{i=1}^{n} \mathbf{m}^{ATT}\left(Z_i, \bar{\theta}^{ATT}\right) \mathbf{m}^{ATT}\left(Z_i, \bar{\theta}^{ATT}\right)' - \Sigma^{ATT} \right\| = o_{P_{n,\delta}}(1),$$

(v) $\frac{1}{\sqrt{n}}\sum_{i=1}^{n} \mathbf{m}^{ATE}\left(Z_i, \theta_n^{ATE}\right) \overset{P_{n,\delta}}{\rightsquigarrow} \mathcal{N}\left(0, \Sigma^{ATE}\right)$ *and* $\frac{1}{\sqrt{n}}\sum_{i=1}^{n} \mathbf{m}^{ATT}\left(Z_i, \theta_n^{ATT}\right) \overset{P_{n,\delta}}{\rightsquigarrow} \mathcal{N}\left(0, \Sigma^{ATT}\right)$.

Proof. Since a proof for the ATT case is similar to the case of ATE, we only focus on proving the claims of the ATE case for the sake of brevity. To prove (i), we first show that under the given assumptions, $\|\hat{\gamma} - \gamma_n\| = o_{P_{n,\delta}}(1)$ holds. Let $l(Z_i, \gamma)$ be the one-observation likelihood for γ in the largest model and $l_n(\gamma) = n^{-1}\sum_{i=1}^{n} l(Z_i, \gamma)$. To establish the uniform weak consistency of the sample likelihood function along $\{P_{n,\delta}\}$, i.e., $\sup_{\gamma\in\Gamma}\left|l_n(\gamma) - E_{P_{n,\delta}}\left[l(Z,\gamma)\right]\right| = o_{P_{n,\delta}}(1)$, consider the mean value expansion of $l(Z,\gamma)$ in γ and bounding from above the absolute derivative term by a parameter-free envelope,

$$\begin{aligned} |l(Z,\gamma) - l(Z,\tilde{\gamma})| \leq & \tilde{F}(W)\|\gamma - \tilde{\gamma}\| \quad \text{for all } \gamma, \tilde{\gamma} \in \Gamma, \text{ where} \\ & \tilde{F}(W) = \left\{ \sup_{\gamma\in\Gamma} \frac{g(W'\gamma)}{G(W'\gamma)} + \sup_{\gamma\in\Gamma} \frac{g(W'\gamma)}{1 - G(W'\gamma)} \right\} \|W\| \end{aligned} \tag{A.3}$$

Compactness of Γ and Assumption REG (iii) then imply that $F(W) \equiv \tilde{F}(W) diam(\Gamma)$ is an integrable envelope of the class of functions $\mathcal{F} = \{|l(\cdot,\gamma) - l(\cdot,\tilde{\gamma})| : \gamma \in \Gamma\}$ with a fixed $\tilde{\gamma}$ with respect to the $L_1(P_0)$-norm. Following the argument of Example 19.7 of van der Vaart (1998) and using the fact that the covering number of a class of functions with radius r is bounded from above by the bracketing number with radius $2r$, the covering number of $\mathcal{F}$ is bounded from above by

$$N(\epsilon\|F\|, \mathcal{F}, \|\cdot\|) \leq \kappa\left[\frac{1}{2\epsilon}\right]^{K+2} < \infty,$$

for every $\epsilon > 0$ and arbitrary semi-norm $\|\cdot\|$ defined on $\mathcal{F}$, where κ is a constant that depends on K and Γ. This leads to the bounded entropy number condition for $\mathcal{F}$. Since $F(W)$ is integrable uniformly over $\{P_{n,\delta}\}$, i.e., $E_{P_{n,\delta}}[F(W)] = E_{P_0}[F(W)] < \infty$ for any $\{P_{n,\delta}\}$ by its construction, Theorem 2.8.1 of van der Vaart and Wellner (1996) yields the desired uniform law of large numbers,

$$\sup_{\gamma\in\Gamma} \left| l_n(\gamma) - E_{P_{n,\delta}}[l(Z,\gamma)] \right| = o_{P_{n,\delta}}(1). \tag{A.4}$$

Combined with compactness of Γ, continuity of $E_{P_{n,\delta}}[l(Z,\gamma)]$ in γ (implied by Assumption DGP (ii)), and the global identification assumption about γ_n (Assumption REG (ii)), Theorem 2.1 of Newey and McFadden (1994) leads to $\|\hat{\gamma} - \gamma_n\| = o_{P_{n,\delta}}(1)$.[19]

The estimator for (μ, τ^{ATE}) in the largest model is

$$\hat{\mu} = \left(\frac{1}{n}\sum_{i=1}^{n} \frac{1-D_i}{1-G(W_i'\hat{\gamma})}\right)^{-1} \left(\frac{1}{n}\sum_{i=1}^{n} \frac{(1-D_i)Y_i}{1-G(W_i'\hat{\gamma})}\right), \quad \hat{\tau}^{ATE} = \left(\frac{1}{n}\sum_{i=1}^{n} \frac{D_i}{G(W_i'\hat{\gamma})}\right)^{-1} \left(\frac{1}{n}\sum_{i=1}^{n} \frac{D_iY_i}{G(W_i'\hat{\gamma})}\right) - \hat{\mu}.$$

Given $\|\hat{\gamma} - \gamma_n\| = o_{P_{n,\delta}}(1)$, we apply Lemma A.1 to the sample averages in the numerator and denominator of $\hat{\mu}$ separately. For a converging sequence $\{\epsilon_n\}$ such that $P_{n,\delta}(\|\hat{\gamma} - \gamma_n\| \le \epsilon_n) \to 1$, let $\Delta_n(Z) = \sup_{\|\gamma-\gamma_n\|\le\epsilon_n} |\frac{1-D}{1-G(W'\gamma)} - \frac{1-D}{1-G(W'\gamma_n)}|$ and $\bar{a}(W) \equiv \sup_{\gamma\in\mathcal{N}} \frac{1}{1-G(W'\gamma)}$, which is by assumption REG (iv), integrable $E_{P_{n,\delta}}(\bar{a}(W)) = E_{P_0}(\bar{a}(W)) < \infty$. For all large n such that $\{\gamma : \|\gamma - \gamma_n\| \le \epsilon_n\} \subset \mathcal{N}$ is true, $E_{P_{n,\delta}}(\Delta_n(Z)) = E_{P_0}\left[(1-G(W'\gamma_n)) \sup_{\|\gamma-\gamma_n\|\le\epsilon_n} |\frac{1}{1-G(W'\gamma)} - \frac{1}{1-G(W'\gamma_n)}|\right]$ implies that the integrand of this expectation is bounded from above by integrable envelope $2\bar{a}(W)$ and converges to zero pointwise at almost all W as $n \to \infty$ by the continuity of $G(\cdot)$. The dominated convergence theorem then implies $E_{P_{n,\delta}}(\Delta_n(Z)) \to 0$ as $n \to \infty$, which validates Condition (i) of Lemma A.1 with $a(Z,\theta) = \frac{1-D}{1-G(W'\gamma)}$. Condition (ii) of Lemma A.1 also holds by Assumption REG (iv). Hence, Lemma A.1 shows $\left|\frac{1}{n}\sum_{i=1}^{n} \frac{1-D_i}{1-G(W_i'\hat{\gamma})} - 1\right| = o_{P_{n,\delta}}(1)$. Following a similar argument, Assumption REG ensures that Conditions (i) and (ii) of Lemma A.1 hold for $a(Z,\theta) = \frac{(1-D)Y}{1-G(W'\gamma)}$, where an integrable envelope can be set at $\bar{a}(Z) \equiv \sup_{\gamma\in\Gamma} \frac{(1-D)Y}{1-G(W'\gamma)}$. We therefore obtain $\left|\frac{1}{n}\sum_{i=1}^{n} \frac{(1-D_i)Y_i}{1-G(W_i'\hat{\gamma})} - E(Y_0)\right| = o_{P_{n,\delta}}(1)$, and by the continuous mapping theorem, $|\hat{\mu} - \mu_0| = o_{P_{n,\delta}}(1)$. A similar argument applied to $\hat{\tau}^{ATE}$ leads to $\left|\hat{\tau}^{ATE} - \tau_0^{ATE}\right| = o_{P_{n,\delta}}(1)$. Hence, $\left\|\hat{\theta}^{ATE} - \theta_n^{ATE}\right\| = o_{P_{n,\delta}}(1)$.

In order to show (ii), it suffices to verify $\left\|\hat{\gamma}^S - \gamma_n\right\| = o_{P_{n,\delta}}(1)$, since stochastic convergence of the rest of parameters in $\hat{\theta}^{ATE,S}$ and $\hat{\theta}^{ATT,S}$ follows by the same argument as in the proof of claim (i) of the current lemma. Consider

$$\left\|\hat{\gamma}^S - \gamma_n\right\| \le \left\|\hat{\gamma}^S - \tilde{\gamma}_n^S\right\| + \left\|\tilde{\gamma}_n^S - \gamma_n^S\right\| + \left\|\gamma_n^S - \gamma_n\right\|. \tag{A.5}$$

[19] Theorem 2.1 of Newey and McFadden (1994) consider fixed DGP asymptotics. Their proof can be adjusted to the case with a drifting sequence of DGPs.

In what follows, we prove each term in the right hand side vanishes asymptotically. By (A.4), the uniform law of large numbers of the sample log likelihood holds also over the constrained parameter space $\Gamma_S = \{\gamma \in \Gamma : \gamma_{S^c} = \mathbf{0}\}$. Hence, combined with the compactness of the parameter space of γ, continuity of the population log-likelihood, and the global identification of $\tilde{\gamma}_n^S$ in the constrained parameter space Γ_S lead to $\left\|\hat{\gamma}^S - \tilde{\gamma}_n^S\right\| = o_{P_{n,\delta}}(1)$ by Theorem 2.1 of Newey and McFadden (1994). Assumptions DGP (iii) implies that the third term in the right hand side of (A.5) is $o(1)$. We show by contradiction that the second term in the right hand side of (A.5) is $o(1)$. Suppose for some $\epsilon > 0$, $\left\|\tilde{\gamma}_n^S - \gamma_n^S\right\| > \epsilon$ holds for all large n. Since $\left\|\gamma_n^S - \gamma_n\right\| = o(1)$ and $E_{P_{n,\delta}}[l(Z,\gamma)]$ is continuous in γ, it holds $E_{P_{n,\delta}}[l(Z,\gamma_n^S)] = E_{P_{n,\delta}}[l(Z,\gamma_n)] + o(1)$. Note that both γ_n^S and $\tilde{\gamma}_n^S$ belong to Γ_S, and hence $E_{P_{n,\delta}}[l(Z,\gamma_n^S)] = E_{P_{n,\delta}}[l(Z,\gamma_n)] + o(1)$ and $\left\|\tilde{\gamma}_n^S - \gamma_n^S\right\| > \epsilon$ contradict the global identification assumption of $\tilde{\gamma}_n^S$ (Assumption REG (ii)). We hence conclude $\left\|\tilde{\gamma}_n^S - \gamma_n^S\right\| = o(1)$.

To show (iii), consider the derivative matrix of the ATE moment conditions,

$$\frac{\partial}{\partial\theta'}\mathbf{m}^{ATE}\left(Z,\theta^{ATE}\right) = \begin{pmatrix} \frac{-g^2+(D-G)(g'-g^2+2g^2G)}{[G(1-G)]^2}WW' & \mathbf{0} & \mathbf{0} \\ \left[-\frac{Dg}{G^2} + \frac{(1-D)g}{(1-G)^2}\right](Y - \tau^{ATE}D - \mu)W' & -\frac{D}{G} & -\frac{D}{G} - \frac{1-D}{1-G} \\ -\frac{Dg}{G^2}(Y - \tau^{ATE}D - \mu) & -\frac{D}{G} & -\frac{D}{G} \end{pmatrix},$$

where we omit the argument of $G(W'\gamma)$, $g(W'\gamma)$, and $g'(W'\gamma) \equiv \frac{d}{da}g(a)\big|_{a=W'\gamma}$ and notate them by G, g, and g', respectively. Having obtained $\|\hat{\gamma} - \gamma_n\| = o_{P_{n,\delta}}(1)$, the boundedness of g and g' (Assumption DGP (ii)) and Assumption REG (iv) guarantee that every element in this derivative matrix satisfy the two conditions of Lemma A.1. Hence, by Lemma A.1, we conclude that $\left\|\frac{1}{n}\sum_{i=1}^n \frac{\partial}{\partial\theta'}\mathbf{m}^{ATE}\left(Z_i,\bar{\theta}^{ATE}\right) - E_{P_{n,\delta}}\left[\frac{\partial}{\partial\theta'}\mathbf{m}^{ATE}\left(Z,\theta_n\right)\right]\right\| = o_{P_{n,\delta}}(1)$ holds. The convergence of $E_{P_{n,\delta}}\left[\frac{\partial}{\partial\theta'}\mathbf{m}^{ATE}\left(Z,\theta_n\right)\right]$ to M^{ATE} follows by the continuity of $G(\cdot)$, $g(\cdot)$, and $g'(\cdot)$, and an application of the dominated convergence theorem.

To show (iv) consider,

$$\begin{aligned} &\mathbf{m}^{ATE}\left(Z,\theta^{ATE}\right)\mathbf{m}^{ATE}\left(Z,\theta^{ATE}\right)' \\ &= \begin{pmatrix} \frac{(D-G)^2g^2}{G^2(1-G)^2}WW' & \left[\frac{Dg}{G^2}(Y_1 - E(Y_1)) - \frac{(1-D)g}{(1-G)^2}(Y_0 - E(Y_0))\right]W' & \frac{Dg}{G^2}(Y_1 - E(Y_1))W' \\ \cdot & \frac{D}{G^2}(Y_1 - E(Y_1))^2 + \frac{1-D}{(1-G)^2}(Y_0 - E(Y_0))^2 & \frac{D}{G^2}(Y_1 - E(Y_1))^2 \\ \cdot & \cdot & \frac{D}{G^2}(Y_1 - E(Y_1))^2 \end{pmatrix}. \end{aligned}$$

Bounded $g(\cdot)$ and Assumption REG (iv) guarantee conditions (i) and (ii) of Lemma A.1. Hence, similarly to the proof of (iii), the conclusion is obtained by applying Lemma A.1.

To show (v), note that Assumption REG (iv) implies the Lindeberg condition for the ATE moment conditions. Therefore, the Lindeberg-Feller central limit theorem for a triangular array of random vectors leads to

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

where $\Sigma_n^{ATE} = E_{P_{n,\delta}}\left[\mathbf{m}^{ATE}\left(Z,\theta_n^{ATE}\right)\mathbf{m}^{ATE}\left(Z,\theta_n^{ATE}\right)'\right]$. Since $\Sigma_n^{ATE} \to \Sigma^{ATE}$ as $n \to \infty$, the desired conclusion follows. ■

Proof of Proposition 2.1. (ATE case) The NPW-ATE estimator in submodel S solves

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

By the mean value expansion around θ_n^{ATE}, we have

$$\begin{aligned}
\mathbf{0} &= \Lambda_S \mathbf{m}_n^{ATE}\left(\theta_n^{ATE}\right) + \Lambda_S\left[\frac{\partial}{\partial\theta'}\mathbf{m}_n^{ATE}\left(\theta_*^{ATE}\right)\right]\begin{pmatrix}\hat{\gamma}^S - \gamma_n \\ \hat{\mu}_S - \mu_0 \\ \hat{\tau}_S^{ATE} - \tau_0^{ATE}\end{pmatrix} \\
&= \Lambda_S \mathbf{m}_n^{ATE}\left(\theta_n^{ATE}\right) + \Lambda_S\left[\frac{\partial}{\partial\theta'}\mathbf{m}_n^{ATE}\left(\theta_*^{ATE}\right)\right]\left[\Lambda_S'\begin{pmatrix}\hat{\gamma}_S - \gamma_{n,S} \\ \hat{\mu}_S - \mu_0 \\ \hat{\tau}_S^{ATE} - \tau_0^{ATE}\end{pmatrix} - \Lambda_{S^c}'\begin{pmatrix}\gamma_{n,S^c} - \gamma_{0,S^c} \\ 0 \\ 0\end{pmatrix}\right],
\end{aligned}$$

where θ_*^{ATE} is a convex combination of $\hat{\theta}^{ATE,S}$ and θ_n^{ATE}. 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.2 (ii), $\left\|\theta_*^{ATE} - \theta_n^{ATE}\right\| = o_{P_{n,\delta}}(1)$. Lemma A.2 (iii) then leads to $\frac{\partial}{\partial\theta'}\mathbf{m}_n\left(\theta_*^{ATE}\right) - M^{ATE} = o_{P_{n,\delta}}(1)$. By Lemma A.2 (v) and Assumption DGP (iii), the asymptotic distribution of $\sqrt{n}\begin{pmatrix}\hat{\gamma}_S - \gamma_{n,S} \\ \hat{\mu}_S - \mu_0 \\ \hat{\tau}_S^{ATE} - \tau_0^{ATE}\end{pmatrix}$ is obtained as

$$\begin{aligned}
&\sqrt{n}\begin{pmatrix}\hat{\gamma}_S - \gamma_{n,S} \\ \hat{\mu}_S - \mu_0 \\ \hat{\tau}_S^{ATE} - \tau_0^{ATE}\end{pmatrix} \\
&= -\left(\Lambda_S M^{ATE}\Lambda_S'\right)^{-1}\Lambda_S\left(\sqrt{n}\mathbf{m}_n^{ATE}\left(\theta_n^{ATE}\right)\right) + \left(\Lambda_S M^{ATE}\Lambda_S'\right)^{-1}\Lambda_S M^{ATE}\Lambda_{S^c}'\begin{pmatrix}\delta_{S^c} \\ 0 \\ 0\end{pmatrix} + o_{P_{n,\delta}}(1) \\
&\overset{P_{n,\delta}}{\rightsquigarrow} -\left(\Lambda_S M^{ATE}\Lambda_S'\right)^{-1}\Lambda_S \times \mathcal{N}\left(0,\Sigma^{ATE}\right) + \left(\Lambda_S M^{ATE}\Lambda_S'\right)^{-1}\Lambda_S M^{ATE}\Lambda_{S^c}'\begin{pmatrix}\delta_{S^c} \\ 0 \\ 0\end{pmatrix} \qquad \text{(A.6)}
\end{aligned}$$

In order to compute the asymptotic variance of $\sqrt{n}\left(\hat{\tau}_S^{ATE} - \tau_0^{ATE}\right)$, we focus on the variance of the bottom element of $-\left(\Lambda_S M\Lambda_S'\right)^{-1}\Lambda_S\mathbf{m}^{ATE}\left(Z_i,\theta_0^{ATE}\right)$. The expectation of the derivative matrix

of the full moment conditions at P_0 is given by

$$\begin{aligned} M^{ATE} &= E_{P_0}\left(\frac{\partial}{\partial\theta'}\mathbf{m}^{ATE}\left(Z_i,\theta_0^{ATE}\right)\right) \\ &= \begin{pmatrix} -E_{P_0}\left(hh'\right) & \mathbf{0} & \mathbf{0} \\ E_{P_0}\left[\left(-\frac{g}{G}\left(\mu_1\left(X\right)-\tau_0^{ATE}-\mu_0\right)+\frac{g}{1-G}\left(\mu_0\left(X\right)-\mu_0\right)\right)W'\right] & -2 & -1 \\ \mathbf{0}' & -1 & -1 \end{pmatrix}. \end{aligned}$$

Hence,

$$\Lambda_S M^{ATE}\Lambda_S' = \begin{pmatrix} -E_{P_0}\left(h_S h_S'\right) & \mathbf{0} & \mathbf{0} \\ E_{P_0}\left[\left(-\frac{g}{G}\left(\mu_1\left(X\right)-\tau_0^{ATE}-\mu_0\right)+\frac{g}{1-G}\left(\mu_0\left(X\right)-\mu_0\right)\right)W_S'\right] & -2 & -1 \\ \mathbf{0}' & -1 & -1 \end{pmatrix},$$

$$\left(\Lambda_S M^{ATE}\Lambda_S'\right)^{-1} = \begin{pmatrix} -E_{P_0}\left(h_S h_S'\right)^{-1} & \mathbf{0} & \mathbf{0} \\ -E_{P_0}\left(\frac{g}{1-G}\left(\mu_0\left(X\right)-\mu_0\right)W_S'\right)E_{P_0}\left(h_S h_S'\right)^{-1} & -1 & 1 \\ E_{P_0}\left(\left(\frac{g}{G}\left(\mu_1\left(X\right)-\tau_0^{ATE}-\mu_0\right)+\frac{g}{1-G}\left(\mu_0\left(X\right)-\mu_0\right)\right)W_S'\right)E_{P_0}\left(h_S h_S'\right)^{-1} & 1 & -2 \end{pmatrix}.$$

By noting

$$\begin{aligned} E_{P_0}\left(\frac{g}{G}\left(\mu_1\left(X\right)-\tau_0^{ATE}-\mu_0\right)W_S'\right) &= E_{P_0}\left(\frac{D-G}{G}\left(\mu_1\left(X\right)-\tau_0^{ATE}-\mu_0\right)h_S'\right) \text{ and} \\ E_{P_0}\left(\frac{g}{1-G}\left(\mu_0\left(X\right)-\mu_0\right)W_S'\right) &= E_{P_0}\left(\frac{D-G}{1-G}\left(\mu_0\left(X\right)-\mu_0\right)h_S'\right), \end{aligned}$$

we can express the bottom element of $-\left(\Lambda_S M\Lambda_S'\right)^{-1}\Lambda_S\mathbf{m}^{ATE}\left(Z_i,\theta_0^{ATE}\right)$ as

$$\begin{aligned} &-E_{P_0}\left[\left(\frac{D-G}{G}\left(\mu_1\left(X\right)-\tau_0^{ATE}-\mu_0\right)+\frac{D-G}{1-G}(\mu_0(X)-\mu_0)\right)h_S'\right]E_{P_0}\left(h_S h_S'\right)^{-1}h_{S,i} \\ &-\left(\frac{D_i}{G_i}+\frac{1-D_i}{1-G_i}\right)\left(Y_i-\tau_0^{ATE}D_i-\mu_0\right) \\ &+2\left(\frac{D_i}{G_i}+\frac{1-D_i}{1-G_i}\right)\left(Y_i-\tau D_i-\mu_0\right)D_i \\ =&L^{\perp}\left\{\left(\frac{D-G}{G}\left(\mu_1\left(X\right)-\tau_0^{ATE}-\mu_0\right)+\frac{D-G}{1-G}(\mu_0(X)-\mu_0)\right)\middle| h_S\right\}+\frac{D_i}{G_i}\left(Y_{1i}-\mu_1\left(X_i\right)\right) \\ &-\left(\frac{1-D_i}{1-G_i}\right)\left(Y_{0i}-\mu_0(X_i)\right)+\left(\Delta\mu(X_i)-\tau_0^{ATE}\right). \end{aligned}$$

These five terms are mean zero and mutually uncorrelated. The sum of their variances therefore gives the asymptotic variance of $\sqrt{n}\left(\hat{\tau}_S^{ATE}-\tau_0^{ATE}\right)$.

Regarding the bias term, (A.6) shows that it is given by the bottom element of the second term

in the right hand side, which is calculated as

$$
\begin{aligned}
&- E_{P_0}\left[\left(\frac{D-G}{G}\left(\mu_1(X)-\tau_0^{ATE}-\mu_0\right)+\frac{D-G}{1-G}(\mu_0(X)-\mu_0)\right)h_S'\right]E_{P_0}\left(h_S h_S'\right)^{-1}E_{P_0}\left(h_S h_{S^c}'\right)\delta_{S^c}\\
&+ E_{P_0}\left[\left(\frac{D-G}{G}\left(\mu_1(X)-\tau_0^{ATE}-\mu_0\right)+\frac{D-G}{1-G}(\mu_0(X)-\mu_0)\right)h_{S^c}'\right]\delta_{S^c}\\
=&E_{P_0}\left[L^{\perp}\left\{\left(\frac{D-G}{G}\left(\mu_1(X)-\tau_0^{ATE}-\mu_0\right)+\frac{D-G}{1-G}(\mu_0(X)-\mu_0)\right)\middle| h_S\right\}h_{S^c}'\right]\delta_{S^c}.
\end{aligned}
$$

(ATT case) The asymptotic distribution of the NPW-ATT estimator follows by replacing M^{ATE} and Σ^{ATE} in (A.6) with M^{ATT} and Σ^{ATT}. Since

$$
M^{ATT}=\begin{pmatrix} -E_{P_0}(hh') & \mathbf{0} & \mathbf{0}\\ E_{P_0}\left(\frac{g}{1-G}\left(\mu_0(X)-\alpha_0\right)W'\right) & -2Q & -Q\\ \mathbf{0}' & -Q & -Q\end{pmatrix},
$$

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

$$
\left(\Lambda_S M^{ATT}\Lambda_S'\right)^{-1}=\begin{pmatrix} -E_{P_0}\left(h_S h_S'\right)^{-1} & \mathbf{0} & \mathbf{0}\\ -\frac{1}{Q}E_{P_0}\left(\frac{g}{1-G}\left(\mu_0(X)-\alpha_0\right)W_S'\right)E_{P_0}\left(h_S h_S'\right)^{-1} & -Q^{-1} & Q^{-1}\\ \frac{1}{Q}E_{P_0}\left(\frac{g}{1-G}\left(\mu_0(X)-\alpha_0\right)W_S'\right)E_{P_0}\left(h_S h_S'\right)^{-1} & Q^{-1} & -2Q^{-1}\end{pmatrix}.
$$

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

$$
\begin{aligned}
&-\frac{1}{Q}E_{P_0}\left(\frac{D-G}{1-G}\left(\mu_0(X)-\alpha_0\right)h_S'\right)E_{P_0}\left(h_S h_S'\right)^{-1}h_{S,i}\\
&-\frac{1}{Q}\left[D_i+(1-D_i)\left(\frac{G_i}{1-G_i}\right)\right](Y_i-\tau D_i-\alpha_0)\\
&+\frac{2}{Q}\left[D_i+(1-D_i)\left(\frac{G_i}{1-G_i}\right)\right](Y_i-\tau D_i-\alpha_0)D_i\\
=&-\frac{1}{Q}L\left\{\left(\frac{D-G}{1-G}\right)\left[\mu_0(X)-\alpha_0\right]\middle| h_S\right\}+\frac{D_i}{Q}\left(Y_i-\mu_1(X)\right)-\frac{1-D_i}{Q}\frac{G_i}{1-G_i}\left(Y_i-\mu_0(X)\right)\\
&+\left(\frac{D-G}{Q}\right)\left[\mu_1(X)-\alpha_0+\frac{G}{1-G}\left(\mu_0(X)-\alpha_0\right)-\tau_0\right]+\frac{G}{Q}\left(\Delta\mu(X)-\tau_0\right).
\end{aligned}
\tag{A.7}
$$

The first term of (A.7) admits the following decomposition,

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

Hence, we can express (A.7) as

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

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

To compute the bias term, focusing on the bottom element of $\left(\Lambda_S M^{ATT}\Lambda_S'\right)^{-1}\Lambda_S M^{ATT}\Lambda_{S^c}'\begin{pmatrix}\delta_{S^c}\\0\\0\end{pmatrix}$ leads to

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

■

The next lemma proves the representation of the Bayes asymptotic MSE (3.6) given in the main text.

Lemma A.3 *Suppose Assumptions DGP and REG. Let $(\hat{B},\hat{\Omega})$ be consistent estimators for (B,Ω) along $\{P_{n,\delta}\}$. For any $\hat{\mathbf{c}}\in\mathcal{C}(B,\Omega)$, the Bayes asymptotic MSE can be represented as (3.6) in the main text.*

Proof. Fix $\delta_{\underline{S}^c}$. Since $(\hat{B},\hat{\Omega})\overset{P_{n,\delta}}{\to}(B,\Omega)$ by the assumption and $\hat{\delta}_{\underline{S}^c}\overset{P_{n,\delta}}{\rightsquigarrow}\Delta_{\underline{S}^c}$, for any $\hat{\mathbf{c}}\in\mathcal{C}(B,\Omega)$, $\hat{\mathbf{c}}=\mathbf{c}\left(\hat{\delta}_{\underline{S}^c},\hat{B},\hat{\Omega}\right)\overset{P_{n,\delta}}{\rightsquigarrow}\mathbf{c}(\Delta_{\underline{S}^c})$ holds by the continuous mapping theorem. Combined with the weak convergence of $\hat{\mathbf{t}}\overset{P_{n,\delta}}{\rightsquigarrow}Z_\tau$ (see equation (3.4) in the main text), the asymptotic MSE can be written as

$$\begin{aligned}
R_\infty(\hat{\mathbf{c}},\delta_{\underline{S}^c})&=\lim_{\zeta\to\infty}E_{\Delta_{\underline{S}^c},Z_\tau|\delta_{\underline{S}^c}}\left[\min\left\{(\mathbf{c}(\Delta_{\underline{S}^c})'Z_\tau)^2,\zeta\right\}\right]\\
&=E_{\Delta_{\underline{S}^c}|\delta_{\underline{S}^c}}\left[\mathbf{c}(\Delta_{\underline{S}^c})'E_{Z_\tau|\Delta_{\underline{S}^c},\delta_{\underline{S}^c}}\left(Z_\tau Z_\tau'\right)\mathbf{c}(\Delta_{\underline{S}^c})\right].
\end{aligned}$$

The claim follows by noting

$$\begin{aligned} E_{Z_\tau|\Delta_{\underline{S}^c},\delta_{\underline{S}^c}}\left(Z_\tau Z_\tau'\right) =& \left[B\delta_{\underline{S}^c}+\Omega_{21}\Omega_{11}^{-1}\left(\Delta_{\underline{S}^c}-\delta_{\underline{S}^c}\right)\right]\left[B\delta_{\underline{S}^c}+\Omega_{21}\Omega_{11}^{-1}\left(\Delta_{\underline{S}^c}-\delta_{\underline{S}^c}\right)\right]' \\ &+\left(\Omega_{22}-\Omega_{21}\Omega_{11}^{-1}\Omega_{12}\right) \\ =& K\left(\Delta_{\underline{S}^c},\delta_{\underline{S}^c}\right). \end{aligned}$$

■

Proof of Proposition 3.1. (i) Solving the Bayes optimal $\hat{\mathbf{c}}(\cdot)$ with risk criterion (3.6) is equivalent to solving for the posterior Bayes action $\hat{\mathbf{c}}(\Delta_{\underline{S}^c})$ for every possible realization of $\Delta_{\underline{S}^c}$. Hence, let $\Delta_{\underline{S}^c}$ be given, and consider minimizing the posterior risk for $\mathbf{c}\left(\Delta_{\underline{S}^c}\right)$ subject to the normalization constraint,

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

If $K_{post}\left(\Delta_{\underline{S}^c}\right)=E_{\delta_{\underline{S}^c}|\Delta_{\underline{S}^c}}\left[K\left(\Delta_{\underline{S}^c},\delta_{\underline{S}^c}\right)\right]$ is nonsingular, this is a quadratic minimization problem with a strictly convex objective function. It therefore has a unique solution and the standard Lagrangian optimization procedure yields $\mathbf{c}^*\left(\Delta_{\underline{S}^c}\right)$ of the proposition. Note that with proper $\mu\left(\delta_{\underline{S}^c}\right)$, the minimized Bayes asymptotic MSE is bounded, because by considering a weight vector that assigns 1 to the largest model, we have

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

where $\omega_{largest}^2$ is the asymptotic variance of the NPW-ATT estimator in the largest model.

(ii) Let $\phi(\cdot:\delta_{\underline{S}^c},\hat{\Omega}_{11})$ be the probability density function of the multivariate normal distribution with mean $\delta_{\underline{S}^c}$ and covariance matrix $\hat{\Omega}_{11}$. Note that $\hat{K}_{post}(\hat{\delta}_{\underline{S}^c})$ can be written as

$$\hat{K}_{post}(\hat{\delta}_{\underline{S}^c})=\frac{\int_{\delta_{\underline{S}^c}}\hat{K}(\hat{\delta}_{\underline{S}^c},\delta_{\underline{S}^c})\phi(\hat{\delta}_{\underline{S}^c}:\delta_{\underline{S}^c},\hat{\Omega}_{11})d\mu(\delta_{\underline{S}^c})}{\int_{\delta_{\underline{S}^c}}\phi(\hat{\delta}_{\underline{S}^c}:\delta_{\underline{S}^c},\hat{\Omega}_{11})d\mu(\delta_{\underline{S}^c})}.$$

By (3.7), $\hat{K}(\hat{\delta}_{\underline{S}^c},\delta_{\underline{S}^c})$ is continuous in $\hat{\delta}_{\underline{S}^c}$ and $\hat{\Omega}_{11}$ in the neighborhood of true Ω_{11}. The Gaussian probability density function $\phi(\hat{\delta}_{\underline{S}^c}:\delta_{\underline{S}^c},\hat{\Omega}_{11})$ is also continuous in $\hat{\delta}_{\underline{S}^c}$ and $\hat{\Omega}_{11}$ in the neighborhood of true Ω_{11}. The dominating convergence theorem then shows that $\hat{K}_{post}(\hat{\delta}_{\underline{S}^c})$ is continuous in $\hat{\delta}_{\underline{S}^c}$ and $\hat{\Omega}_{11}$ in the neighborhood of true Ω_{11}. Therefore $\hat{K}_{post}(\hat{\delta}_{\underline{S}^c})\overset{P_{n,\delta}}{\rightsquigarrow}K_{post}(\Delta_{\underline{S}^c})$ and $\mathbf{c}^*(\hat{\delta}_{\underline{S}^c},\hat{B},\hat{\Omega})\overset{P_{n,\delta}}{\rightsquigarrow}\mathbf{c}^*(\Delta_{\underline{S}^c})$ hold by the continuous mapping theorem. Hence, $\mathbf{c}^*\left(\hat{\delta}_{\underline{S}^c},\hat{B},\hat{\Omega}\right)$ attains the lower bound of the Bayes asymptotic MSE, $R_\infty^{Bayes}\left(\mathbf{c}^*\left(\hat{\delta}_{\underline{S}^c},\hat{B},\hat{\Omega}\right)\right)=\inf_{\hat{\mathbf{c}}\in\mathcal{C}(B,\Omega)}R_\infty^{Bayes}\left(\hat{\mathbf{c}}\right)$. ■

Proof of asymptotic validity of $CI_{1-\beta}^{ATT}(\hat{\delta}_{\underline{S}^c},\hat{\mathbf{t}})$.

Let $\delta_{\underline{S}^c}$ be given. By the construction of $CI^{ATT}_{1-\beta_1}(\cdot,\cdot|\delta_{\underline{S}^c})$ and the weak convergence of $(\hat{\delta}_{\underline{S}^c}, \hat{\mathbf{t}})$ shown in ((3.4)), it holds

$$\begin{aligned}
1-\beta_1 &= \lim_{n\to\infty} P_{n,\delta}\left(\tau_n^{ATT} \in CI^{ATT}_{1-\beta_1}(\hat{\delta}_{\underline{S}^c}, \hat{\mathbf{t}}|\delta_{\underline{S}^c})\right) \\
&\leq \lim_{n\to\infty} P_{n,\delta}\left(\tau_n^{ATT} \in CI^{ATT}_{1-\beta_1}(\hat{\delta}_{\underline{S}^c}, \hat{\mathbf{t}}|\delta_{\underline{S}^c}), \delta_{\underline{S}^c} \in CS_{1-\beta_2}\right) + \lim_{n\to\infty} P_{n,\delta}\left(\delta_{\underline{S}^c} \notin CS_{1-\beta_2}\right) \\
&\leq \lim_{n\to\infty} P_{n,\delta}\left(\tau_n^{ATT} \in CI^{ATT}_{1-\beta}(\hat{\delta}_{\underline{S}^c}, \hat{\mathbf{t}})\right) + \beta_2
\end{aligned}$$

where the third line follows by noting that on the event $\delta_{\underline{S}^c} \in CS_{1-\beta_2}$, the union confidence intervals $CI^{ATT}_{1-\beta}(\hat{\delta}_{\underline{S}^c}, \hat{\mathbf{t}})$ contain $CI^{ATT}_{1-\beta_1}(\hat{\delta}_{\underline{S}^c}, \hat{\mathbf{t}}|\delta_{\underline{S}^c})$. Hence, $\lim_{n\to\infty} P_{n,\delta}\left(\tau_n^{ATT} \in CI^{ATT}_{1-\beta}(\hat{\delta}_{\underline{S}^c}, \hat{\mathbf{t}})\right) \geq 1-\beta_1-\beta_2 = 1-\beta$. The valid coverage does not depend on the value of δ nor a construction of the sequences $\{P_{n,\delta}\}$, and thereby the asymptotic coverage is uniformly valid over the class of DGPs satisfying Assumptions DGP (i) - (ii) and REG. ■

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 (2014), "New Evidence on the Finite Sample Properties of Propensity Score Reweighting and Matching Estimators," *Review of Economics and Statistics*, 96(5), 885–897.

[4] Carrasco, M., J.P. Florens, and E. Renault (2007). "Linear Inverse Problems in Structural Econometrics Estimation Based on Spectral Decomposition and Regularization," *Handbook of Econometrics, Vol VI*, J.J. Heckman and E. Leamer (Eds.), Elsevier, 5633-5751.

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

[6] Cheng, X., Z. Liao, and R. Shi (2015), "Uniform Asymptotic Risk of Averaging GMM Estimator Robust to Misspecification," *unpublished manuscript*, University of Pennsylvania and UCLA.

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

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

[9] 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.

[10] 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.

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

[12] Farrell, M.H. (2013), "Robust Inference on Average Treatment Effects with Possibly More Covariates than Observations," *unpublished manuscript*, University of Michigan.

[13] 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.

[14] 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.

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

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

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

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

[19] Hansen, B.E. (2014a), "Model Averaging, Asymptotic Risk, and Regressor Groups," *Quantitative Economics*, 5, 495-530.

[20] Hansen, B.E. (2014b), "Efficient Shrinkage in Parametric Models," *unpublished manuscript*, University of Wisconsin, Madison.

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

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

[23] 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.

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

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

[26] 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.

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

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

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

[30] 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.

[31] James, W. and C.M. Stein (1961), "Estimation with Quadratic Loss," *Proceedings of the Fourth Berkeley Symposium on Mathematical Statistics and Probability*, 1, 361-380.

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

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

[34] Lehmann, E.L. and J.P. Romano (2005), *Testing Statistical Hypotheses, Third Edition.* Springer, New York.

[35] Liu, C-A., "Distribution Theory of the Least Squares Averaging Estimator," *unpublished manuscript*, National University of Singapore.

[36] Liu, Q. and R. Okui (forthcoming), "Heteroskedasticity-robust C_p Model Averaging," *Econometrics Journal.*

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

[38] 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.

[39] Millimet, D.L. and R. Tchernis (2009), "On the Specification of Propensity Scores, With Applications to the Analysis of Trade Policies," *Journal of Business Economics and Statistics*, 27, 397-415.

[40] Newey, W.K. and D.L. McFadden (1994), "Large Sample Estimation and Hypothesis Testing," *Handbook of Econometrics, Vol IV*, R.F. Engle and D.L. McFadden eds., 2113-2245. Elsevier, New York.

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

[42] Song, K. (2014), "Point Decisions for Interval Identified Parameters," *Econometric Theory*, 30, 334-356.

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

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

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

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

[47] 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.

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

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

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