Covariate Selection and Model Averaging in Semiparametric Estimation of Treatment Effects

Toru Kitagawa (University College London),
Chris Muris (Simon Fraser University)

01. Setting

- A **random sample** of measurements on individuals is available
- Some individuals were affected by a **program**
- **Selection** into treatment on **observables**
- Focus: average treatment effect for the treated (**ATT**)
- Estimation using a **propensity-score** weighting estimator
  - these estimators are **very common** in empirical practice
  - simulation evidence suggests **excellent performance** (Busso, DiNardo, McCrary; REStat, 2011)

02. Problem

- **How to choose** the covariates that enter the propensity score?
    1. Which variables X to choose?
    2. Which functions of X to include?
- There may be a **bias-variance tradeoff**
    - leaving out relevant covariates: omitted variable bias
    - including redundant variables: increases variance
    - Intuition: if there are **many** regressors, we may not want to use all of them
- Common practice: "put everything in"
- Which selection of covariates/functional form is **optimal**?

03. Contribution

1. We show that a bias-variance **tradeoff exists**
2. We propose a **data-driven** way of **selecting** regressors for the propensity score **(model selection),**
   - based on minimizing the estimated mean squared error
3. We propose an **optimal** way of averaging over candidate specifications **(model averaging)**
4. Averaging estimator outperforms "put-everything-in" by up to **25**-**30**% (MSE, simulations)

04. Motivation: Treatment effects

- Effect of motherhood on wages
  - Simonsen and Skipper (JAE, 2006)
  - **29027** observations, **172** covariates
- Development project aid money on rural rehabilitation projects
  - van de Walle and Mu (JDE, 2007)
  - **194** observations, **35** covariates
- Effect of CEO awards on firm productivity
  - Malmendier and Tate (QJE, 2009)
  - **71418** observations, **100's** of covariates
- Hirano, Imbens, and Ridder (ECTA, 2003):
  - **series** estimator is efficient
  - In practice, researcher must choose **number of terms**

05. Model: potential outcomes

- $\{(Y_i, D_i, X_i), i = 1, \cdots, n\}$ is a random sample of size n. **treatment indicator** $D_i \in \{0, 1\}$; **scalar outcome** Y_i; vector of **covariates** $X_i = (X_{i1}, \cdots, X_{iL})$;
- Potential outcomes $(Y_i(1), Y_i(0))$, so that
$$Y_i = \begin{cases} Y_i(1) & \text{if } D_i = 1 \\ Y_i(0) & \text{if } D_i = 0 \end{cases}$$
- **Assumption 1: Unconfoundedness.**
$(Y_i(1), Y_i(0)) \perp D_i | X_i$
- **Assumption 2: Propensity score.** For a known vector $W_i \equiv W(X_i) \in \mathbb{R}^K$ of linearly independent functions of X_i, there exist a unique $\gamma_0 \in \mathbb{R}^K$ such that
$P(D_i = 1 | X_i) = G\left(W_i' \gamma_0\right)$ for a known link function $G(\cdot)$
- **Assumption 3: Strict overlap.** There exists an $\epsilon > 0$ such that $G\left(W(x)' \gamma_0\right) \leq \epsilon < 1$ for all values of $x \in \text{supp}(X_i)$

06. Estimation: Normalized propensity weights

- Step 1: Estimate propensity score parameter $\hat{\gamma}$, by ML
- Step 2:

$$\hat{\tau}_{NPW} = \frac{1}{n}\sum_{i=1}^{n}\left[\frac{D_i Y_i}{\sum_{i=1}^{n} D_i/n} - \frac{\frac{G\left(W_i'\hat{\gamma}\right)(1-D_i)}{\left(1-G\left(W_i'\hat{\gamma}\right)\right)} Y_i}{\sum_{i=1}^{n} \frac{G\left(W_i'\hat{\gamma}\right)(1-D_i)}{\left(1-G\left(W_i'\hat{\gamma}\right)\right)}/n}\right]$$

- **Alternatively**, use only a **subset** of covariates $W_{S,i} \subset W_i$

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

- Leads to **collection** of estimators

$$\{\hat{\tau}_{S,NPW}\}_S$$

07. Local misspecification: motivation

- **Standard asymptotics:** no bias-variance tradeoff
- For the **full** model estimator, $\sqrt{n}\left(\hat{\tau}_{NPW} - \tau_0\right) \to \mathcal{N}\left(0, \omega^2_{NPW}\right)$
- For any **subset estimator**, $\sqrt{n}\left(\hat{\tau}_S - \tau_S\right) \to \mathcal{N}\left(0, \omega^2_S\right)$, where $\tau_S = \text{plim}\left(\hat{\tau}_S\right) \neq \tau_0$
- Then

$$
\begin{aligned}
\sqrt{n}\left(\hat{\tau}_S - \tau_0\right) &= \sqrt{n}\left(\hat{\tau}_S - \tau_S\right) + \sqrt{n}\left(\tau_S - \tau_0\right) \\
&\to \mathcal{N}\left(0, \omega^2_S\right) \pm \infty
\end{aligned}
$$

- Asymptotically, we **always** prefer the big model. **Counterintuitive.**

08. Asymptotic distribution $\hat{\tau}_S$

$$
\begin{aligned}
\sqrt{n}(\hat{\tau}_S - \tau_0) &\rightarrow \mathcal{N}(0, \omega_S^2) + \text{bias}_S, \\
\omega_S^2 &= \frac{1}{Q^2} E\left[L^2 \left((D_i - G_i) \frac{1 - 2G_i}{1 - G_i} (\mu_0(X_i) - \alpha_0) \middle| h_S \right) \right] + \\
\text{bias}_S &= \frac{1}{Q} E\left[L^{\perp} \left(\frac{D_i - G_i}{1 - G_i} (\mu_0(X_i) - \alpha_0) \middle| h_S \right) h'_{S^c} \right] \delta_{S^c}
\end{aligned}
$$

- $\omega_S^2 \nearrow$ as $\dim(W_{S,i}) \nearrow$
- $|\text{bias}_S|$ tends to $\searrow$ as $|\delta_{S^c}| \searrow$ or $\dim(W_{S,i}) \nearrow$.

- Notation
  - $G_i = G\left(W_i' \gamma_0\right)$, $\mu_0(X_i) = E(Y_i(0) | X_i)$, $\alpha_0 = E(Y_i(0) | D_i = 1)$
  - $h_S = \frac{(D_i - G_i) G_i}{G_i(1 - G_i)} W_{S,i}$ and $L(\cdot | \cdot)$ is a projection,
  - S^C is an index for the regressors not in S, e.g. $W_{S^C,i} = W_i \backslash W_{S,i}$

09. Model selection: FIC

- We have a collection $\{\hat{\tau}_S\}$ of estimators
- We want to select the estimator with the lowest MSE
- Not **feasible:** MSE must be estimated
- **Focussed information criterion (FIC)** approach (Claeskens/Hjort, JASA, 2003; CUP, 2008):
    - Assume that the **full** model is **correctly** specified
    - **Focus** is on τ_0, γ is a nuisance parameter
    - FIC: an unbiased estimate of $MSE(\hat{\tau}_S)$ for each estimator
    - **Select** estimator with the lowest FIC

10. Model selection: MSE estimation

- Mean squared error for $\hat{\tau}_S$ can be written

$$\text{MSE}_S = \omega_S^2 + b_S' \delta_{S^c} \delta_{S^c}' b_S$$

- Consistent estimators for ω_S^2 and b_S' are available from full model estimation
- **Problem:** No consistent estimator for $\delta = \sqrt{n}\,(\gamma_n - \gamma_0)$
- For example, consider$\hat{\delta} = \sqrt{n}\,(\hat{\gamma} - \gamma_0) \to \mathcal{N}\,(\delta, V)$
- $\hat{\delta}$ is **unbiased** but not consistent
- For MSE, we are interested in $\delta\delta'$. Use: $\widehat{\delta\delta'} = \hat{\delta}\hat{\delta}' - \hat{V}$
- Now, all **ingredients** for FIC model selection are available

Model selection: overview

1. Specify a largest model by choosing W_i
2. Specify which submodels S are considered
3. Obtain the NPW estimator using the full set of covariates
   - Also provides $\hat{\omega}_S^2$, $\hat{b}_S$ for each submodel S, and $\widehat{\delta\delta'}$
4. For each estimator, compute
   $FIC(S) = \widehat{MSE}_S = \hat{\omega}_S^2 + \hat{b}_S' \widehat{\delta_{S^c}\delta_{S^c}'} \hat{b}_S'$
5. Choose the estimator with minimum $FIC(S)$
6. This is the FIC selection estimator for ATT

11. Model averaging

- Model selection estimators are "discontinuous" in $\hat{\delta}$
- An alternative is to consider model averaging estimators

$$\hat{\tau}_{avg} = \sum_S c_S\left(\hat{\delta}\right)\hat{\tau}_S, \sum_S c_S\left(\hat{\delta}\right) = 1$$

- Model selection: $c_S = \begin{cases} 1 & \text{if } FIC(S) \text{ is minimized at } S \\ 0 & \text{if not} \end{cases}$
- Alternative: assign smooth weights, e.g. $c_S = \frac{FIC^{-1}(S)}{\sum_S FIC^{-1}(S)}$

12. Optimal averaging

- The distribution of the averaging estimator is given by

$$\sqrt{n}(\hat{\tau}_{avg} - \tau_n) = \sum_{S} c_S\left(\hat{\delta}\right)\sqrt{n}(\hat{\tau}_S - \tau_n)$$

- The MSE converges to

$$MSE(\hat{\tau}_{avg}) \to E_{\hat{\delta}|\delta}\left[c(\hat{\delta})' K(\hat{\delta}, \delta) c(\hat{\delta})\right],$$

with $c\left(\hat{\delta}\right)$ the vector of weights, $\hat{\delta} \sim \mathcal{N}(\delta, \Sigma_\delta)$, and

$$K(\hat{\delta}, \delta) = V + (A_1\delta + A_2\hat{\delta})(A_1\delta + A_2\hat{\delta})'$$

- MSE-minimizer not **feasible**: depends on the true value of δ

13. Optimal averaging: Statistical decision

We propose to use weights that solve

$$c^* \left(\hat{\delta}\right) = \arg\min_{c(\cdot)} \int_{\delta} E_{\hat{\delta}|\delta} \left[c(\hat{\delta})' K(\hat{\delta}, \delta) c(\hat{\delta}) \right] d\mu(\delta)$$

where $\mu(\delta)$ is a prior on δ

Proposition:

Let $\mu(\delta)$ be a proper prior, and assume that $K_{post}(\hat{\delta}) = E_{\delta|\hat{\delta}}(K(\hat{\delta}, \delta))$ is nonsingular. Then

$$c^*(\hat{\delta}) = \frac{1}{\iota' K_{post}(\hat{\delta})\iota} K_{post}(\hat{\delta})\iota.$$

14. Toy model: setup

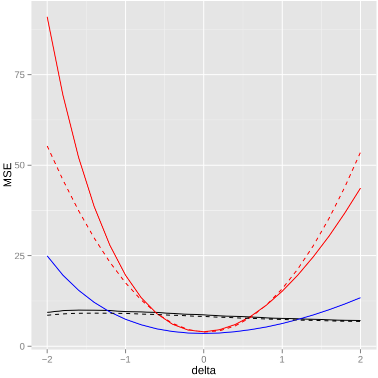
- X_i is binary covariate, with $P(X_i = 1) = 0.5$
- If $X_i = 0$, then $Y_i = 0$
- If $D_i = 1$, then $Y_i = 1$
- $P(D = 1|X = 0) = 0.4$, $P(D = 1|X = 1) = \gamma_1$
- $E(Y_i|D = 0, X = 1) = \mu_Y$ and variance σ_Y^2

In this model, the expression for the ATE and ATT are straightforward:

$$\begin{aligned} ATE &= -p_X \mu_Y \\ ATT &= -P(X = 1|D = 1)\mu_Y \end{aligned}$$

1. Estimator 1: No covariates: $\hat{\tau}_0 = \frac{1}{n}\sum_i D_i Y_i$
2. Estimator 2: Include covariate: $\hat{\tau}_f = -\hat{q}_1 \hat{\mu}_Y$

15. Toy model: results



MSE for the estimator with covariate (black), the estimator without covariate (red), and the averaging estimator (blue). Dotted lines are asymptotic approximations, solid lines are simulation results.

Simulation model

- Model for **simulations**:

$$
\begin{aligned}
P(D_i = 1 | X_i) &= \Lambda(\gamma_0 + X_i'\gamma), \\
Y_i(0) &= \beta_{00} + X_i'\beta_0 + u_{0i}, \\
Y_i(1) &= \beta_{10} + X_i'\beta_1 + u_{1i}, \\
u_{ji} | X_i &\sim N\left(0, \sigma_j^2\right), \, j \in \{0, 1\}, \\
X_i &\sim \mathcal{N}_K\left(0_K \, , \, cI_K + (1-c)\iota_K\iota_K'\right).
\end{aligned}
$$

- **Logit** link, **normality** for regressors and disturbances, **linear** outcome equations
- Note: simulation results **do not depend** on the local misspecification framework

Benchmark values

Parameter	Value	Interpretation
n	300	Moderate sample size
K	4	4 regressors
c	0.7	Corr(X_1, X_2)=0.3
$\gamma_1 = \beta_{11}$	1	X_1 is the important regressors
$\gamma_k,\ k > 1$	0.1	Other regressors are less important
$\beta_{1,k},\ k > 1$	$k/10$	Heterogeneous treatment effects
β_0	0	$Y_i(0) = u_{0i}$
γ_0, β_{10}	1; 1	-
$\sigma_0 = \sigma_1$	0.1	-
Reps	9000	9000 Monte Carlo reps
BS reps	1000	1000 bootstrap reps for $\hat{\Omega}$
		2^{K-1} submodels

Table : Parameter values for the benchmark simulations.

Benchmark results

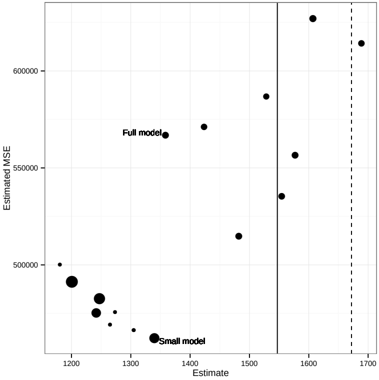
Estimator	All submodels Bias	Var	MSE
$\{X_1\}$	6.29	3.63	4.02
$\{X_1, X_2\}$	4.48	3.80	4.00
$\{X_1, X_2, X_3\}$	3.33	3.97	4.08
$\{X_1, X_2, X_3, X_4\}$	2.50	4.11	4.17
$\vdots$	$\vdots$	$\vdots$	$\vdots$
$\{X_2, X_3, X_4\}$	72.79	1.66	54.64
Best submodel	4.44	3.80	4.00
Selection	4.24	3.72	3.90
Bayes	4.84	2.78	3.01
HC	5.54	3.35	3.66
invFIC	6.88	3.06	3.53
Relative efficiency		72%	

Table : All values were multiplied by 100.

16. Application: National Supported Work Demonstration

- We apply the estimators to Lalonde (AER, 1986) and Dehejia and Wahba (JASA, 1999)
- Effect of a labor market **training** on post-program **earnings**
- **Experimental** results (dotted vertical line): \$1631 (sd: 637)
- Lalonde: results **cannot be replicated** with regression methods and PSID/CPS
- DW: results **can** be replicated, using propensity score methods
- Our results (triangle): more precise, closer to experimental

17. Application: Results



Dots represent individual submodel estimates. The dotted line represents the experimental estimate and its standard error. The **solid line** represents our **averaging** estimate.